Journal of Ethics and Social Philosophy
VOL. 24, NO. 2 · MARCH 2023
https://doi.org/10.26556/jesp.v24i2.2494

IN SEARCH OF A STABLE CONSENSUS

RAWLS'S MODEL OF PUBLIC REASON AND ITS CRITICS

Cyril Hédoin

RAWLS'S political philosophy figures as the most important contribution to what is sometimes called *public reason liberalism*. The concept of public reason, however, makes its appearance only in Rawls's late writings, especially *Political Liberalism* and "The Idea of Public Reason Revisited." The central place that this concept occupies in these writings is associated with Rawls's "political turn" that marks the transition from *A Theory of Justice* to *Political Liberalism*. The recent scholarship on Rawlsian political philosophy has established that this political turn follows Rawls's attempt to overcome the unsatisfactory treatment of the problem of stability of the well-ordered society in *A Theory of Justice*.[1] The contours of Rawls's later solution to this problem are relatively well known. Principles of justice are merely political and should be publicly endorsable by each person within one's "comprehensive doctrine." This establishes an overlapping consensus that guarantees that each member of the society has all-things-considered reasons to support a shared liberal conception of justice.

A key feature of this Rawlsian account of stability is the requirement that the members of the well-ordered society abide by principles of justice for *shared public reasons*. This "stability for the right reasons" requirement has recently been the target of a series of criticisms. The criticisms are mainly internal to public reason liberalism and come from post-Rawlsian scholars who endorse what I shall call a *diversity-convergence account of stability*. They emphasize in particular the inability of public reason to solve the assurance problem with which the members of the well-ordered society are confronted. This paper discusses the debate over Rawls's model of public reason as an account of stability that these criticisms trigger. I suggest a way to preserve the Rawlsian account of stability based on the possibility that members of the well-ordered society can reasonably make use of what I call *community-based reasoning* to solve the assurance problem. This solution is nonetheless fragile in the presence

1 Weithman, *Why Political Liberalism?*

of unreasonable persons. This result calls for a reassessment of the project of public reason liberalism.

The rest of the paper is structured as follows. Section 1 presents the problem of stability in Rawlsian political philosophy. Section 2 characterizes Rawls's model of public reason as an account of stability. Section 3 discusses the various criticisms of Rawls's account of stability developed by proponents of the diversity-convergence account of stability. Section 4 suggests an alternative solution in terms of community-based reasoning that remains compatible with Rawls's model. Section 5 considers the prospects of establishing a well-ordered society based on a minimal set of shared reasons when this last solution fails. Section 6 briefly concludes.

1. The Problem of Stability in Rawlsian Political Philosophy

Rawls's theory of justice is part of the very large family of social contract theories. As are most—if not all—normative accounts belonging to this family, it is confronted with two related problems: the *justificatory problem* and the *stability problem*. The justificatory problem can be formulated as follows: On what basis can a set P of normative principles about what is right and good be justified to some set N of persons? The stability problem is concerned with a different question: What are the conditions, if any, under which the members of N will abide by P?

There are several ways to meet these requirements. Contractualist moral and political theories generally solve the justificatory problem by positing a set N^* of idealized individuals placed under idealized circumstances C^*. The characteristics of N^* and C^* are such that it is claimed that the set of principles P over which the members of N^* will agree (either through a unanimous choice or a compromise through a bargain) have a particular normative significance. The stability problem then surfaces immediately. Even if the set of principles P is appropriately justified, the fact that the nonidealized members of N are put under nonidealized circumstances C does not guarantee that the principles P will be respected and implemented. A tension between the justificatory problem and the stability problem thus appears.[2] There are at least two reasons to impose severe restrictions on N^* and C^*. First, our considered intuitive judgments may lead us to think that only agreements based on specific reasons or motivations and reached under specific circumstances can count as morally relevant. Second, idealization is helpful to restrict the candidate principles potentially belonging

2 See Thrasher, "Agreeing to Disagree," for a similar characterization that emphasizes the tension between what he calls the existence problem and the stability problem.

to *P*. Insufficiently restricted characterizations of *N** and *C** run the risk of leading to complete indeterminacy. But of course, the more *N** and *C** are idealized, the less likely it is that the members of *N* (actual members of society), put under nonidealized circumstances *C*, will actually implement the principles.

As Gaus notes, in Rawls's contractualism, the "members of the justificatory public" belonging to set *N** are idealized in several ways.[3] In particular, the members of *N** are assumed to be good-willed persons. In Rawls's words, they are *reasonable* persons moved by a sense of justice, respecting others as free and equal persons. They are, moreover, appropriately motivated: considerations foreign to justice are temporarily bracketed in their practical deliberations. This idealization of the members of *N** is of course closely related to the particular circumstances under which they are put. The original position, viewed as a "procedure of construction," is constructed in such a way that idealized individuals cannot but rationally ignore considerations that are not related to justice.[4] Rawls's theory of justice thus faces the stability problem outlined above. Once idealized members of the justificatory public have agreed on a set of principles of justice regulating the basic structure of the society, it has still to be established that members of the society will respect the institutions that implement the principles.

Commentators of Rawls's scholarship have recently helped to clarify the nature of the stability problem in Rawlsian political philosophy.[5] Two threats for stability can be identified once the veil of ignorance characterizing the Rawlsian original position is lifted. The first threat refers to the fact that once the veil is lifted, each person will pursue her plans as conceived by her conception of the good life. While members of *N** only take into account considerations of justice based on a "thin" theory of the good, actual members of the society use their full deliberative rationality to pursue their unrestricted conception of the good. This is the problem of "justificatory instability."[6] Weithman, following Rawls, characterizes this problem as an *n*-person prisoner's dilemma.[7] Each member of *N* may be tempted to free ride, meaning that one's conception of the good may rationally encourage one to behave wrongly (i.e., not to follow the justice principles agreed on), no matter how the other members are behaving. Suppose that the problem is solved along the lines Rawls suggests: members of *N* have a "sense of justice" that "leads [them] to promote just schemes and to do [their]

3 Gaus, "A Tale of Two Sets," 305–6.

4 Rawls, "Kantian Constructivism in Moral Theory."

5 Gaus, "A Tale of Two Sets"; Weithman, *Why Political Liberalism?*

6 Gaus, "A Tale of Two Sets," 307–15.

7 Weithman, *Why Political Liberalism?* 48; Rawls, *A Theory of Justice*, 505.

share in them when [they] believe that others, or sufficiently many of them, will do theirs."[8] Here comes the second threat: persons who have a sense of justice sufficiently strong to overcome the temptation to free ride will act rightly if and only if they expect others to do so. This is a mutual assurance problem that generalizes to *n* players the following two-person assurance game (fig. 1).

		Person 2	
		Act rightly	Act wrongly
Person 1	Act rightly	3; 3	0; 2
	Act wrongly	2; 0	1; 1

FIGURE 1 A Two-Person Assurance Game

In such an assurance game, acting rightly is rational as long as person 1 believes that person 2 is sufficiently likely to act rightly. Interactive reasoning further implies that person 1 will have this belief only if she believes that person 2 believes that she is sufficiently likely to act rightly, and so on *ad infinitum*. Solving the assurance problem requires finding the appropriate ground for a full (and thus infinite) hierarchy of beliefs based on which acting rightly is rational.[9]

The stability problem is the topic of part 3 of *A Theory of Justice*. To understand Rawls's proposed solution to it, it is important to acknowledge two requirements that Rawls imposes on any plausible solution. These requirements indicate that Rawls is concerned with a specific kind of stability. First, Rawls's characterization of a well-ordered society points out that any kind of stable state will not do. Only stable *just* states are acceptable—that is, states in which people behave following the principles of justice. Second, principles of justice should be *inherently stable* rather than stable by imposition: "A conception of justice is inherently stable if a society that is well-ordered by it generally maintains itself in a just general equilibrium and is capable of righting itself when that

8 Rawls, *A Theory of Justice*, 236.

9 Generalizing the assurance game to *n* persons leads to two significant differences. First, each player has to form a belief on the likely number of persons who will act rightly. The hierarchy of beliefs is then defined on each person's belief about the number of persons who will act rightly. Second, we may reasonably imagine that different persons have different belief thresholds above which acting rightly is rational. Persons with a more developed sense of justice will act rightly even if they expect a relatively low number of other persons to do so. This makes the dynamic more complicated, but in the end, such an *n*-person assurance game would still have two stable equilibria: one where a low number of persons act rightly and one where a high number of persons act rightly. See Granovetter ("Threshold Models of Collective Behavior") for classical dynamic threshold models of this kind.

equilibrium is disturbed."[10] Rawls's proposed solution to the stability problem in *A Theory of Justice* proceeds through an argument for the *congruence of the right and the good*.[11] In a nutshell, Rawls contends that the institutions of the just society will solve the stability problem if and only if they are able to elicit in the members of *N* a sense of justice that they have sufficiently strong reasons to keep such that acting wrongly would not be rational. Because members of *N* are assumed to rationally pursue what they consider to be good, institutions will foster inherent stability if and only if the sense of justice is part of the conceptions of the good of members of *N*. Rawls claims that this necessary and sufficient condition is fulfilled as soon as we assume that members of *N* want to live up to a small number of ideals (friendship, personal conduct, association) that are constitutive of justice as fairness. The congruence between the right and the good solves both the justificatory instability and the assurance problems at the same time. It establishes that members of *N* have all-things-considered reasons to live up to their sense of justice and thus to abide by the principles agreed on by their ideal counterparts, the members of *N**.[12]

2. RAWLS'S MODEL OF PUBLIC REASON

Specialists of Rawls's scholarship disagree over the reasons for and the meaning of Rawls's political turn. This political turn is fully stated in *Political Liberalism* but was engaged by Rawls in the 1980s, in particular in important essays such as "Justice as Fairness: Political not Metaphysical" and "The Domain of the Political and Overlapping Consensus." Justice as fairness and more generally all plausible liberal principles of justice are reframed as *political* principles. Their reasonability as principles of justice is presented as being independent of metaphysical truths and more generally of the truth value of any proposition belonging to areas of philosophy other than political philosophy. In particular, Rawls's political liberalism is sometimes read as an attempt to reformulate justice as fairness by substituting a *political* conception of persons as free and equal citizens for the metaphysical conception on which *A Theory of Justice* is thought

10 Weithman, *Why Political Liberalism?* 45.

11 See Weithman, *Why Political Liberalism?* esp. chs. 2–4, for a detailed discussion of the congruence argument and its problems.

12 In fact, the congruence argument is stronger than that. Rawls argues that persons in the original position would not agree on principles that they do not expect to be stable in the relevant sense. That means that the congruence argument states that the sense of justice and the related ideals are actually part of the thin theory of the good that is shared by all members of *N**.

to build. As Weithman forcefully argues, this reading is at best misleading.[13] First, it is not clear that the argument for justice as fairness in *A Theory of Justice* depends on a metaphysical conception of the person. Second, and more importantly, it misses the point (stated explicitly by Rawls in several places) that the main problem is the treatment of the stability problem in *A Theory of Justice*.

Rawls became skeptical about the congruence argument as an account of the ability of just political institutions to foster inherent stability because it depends on assumptions that contradict the very nature of a liberal society. Liberal political institutions favor the development of a great diversity of views and opinions. These views and opinions are the product of partial or full "comprehensive doctrines" held by members of *N*. Comprehensive doctrines are constituted by values and trade-offs between these values that express conceptions of the good life. The "burdens of judgments" correspond to the many causes Rawls identifies that explain why, in a liberal society, we cannot avoid a reasonable disagreement over conceptions of the good life. The problem with the congruence argument is that it (apparently) solves the stability problem but at a considerable cost: it disregards the burdens of judgment by assuming that members of *N* all endorse ideals that belong to a specific comprehensive doctrine, justice as fairness. Thus, it fails to establish what it needs to establish: that just and liberal political institutions would generate their own stability. In this sense, justice as fairness needs to be reformulated as a political rather than comprehensive doctrine. The reason, however, is not related to the conception of the person *per se*; rather, this account of justice has to be vindicated without assuming that members of *N* endorse any specific conception of the good. A new solution to the stability problem thus has to be found.

What has to be demonstrated is that in a well-ordered society, members of *N* have reasons to abide by political principles of justice for shared reasons that do not depend on their comprehensive doctrines. If these reasons are sufficiently strong, they will solve the justificatory instability problem. To solve the assurance problem, however, they also have to be *public*. Public reason makes its appearance in Rawls's political account as a coordination device that helps members of *N* obtain the required assurance that everyone will act rightly because everyone expects everyone else to act rightly, all this being commonly known among the members of *N*. To understand this, let me briefly expose a stylized version of what I call *Rawls's Model of Public Reason* (henceforth, RMPR).[14]

13 Weithman, *Why Political Liberalism?*

14 I call it a stylized version because I ignore several details Rawls discusses at length in "The Idea of Public Reason Revisited." Also, as Gaus and Van Shoelandt demonstrate, Rawls's political liberalism has the structure of a baroque cathedral that can be represented by several competing and not entirely consistent models ("Consensus on What?"). My

The RMPR can be seen as an attempt to solve the justificatory and stability problems at the same time. In this model, justification comes first through the agreement over political principles of justice by members of N^* put under circumstances C^*. The original position is a procedure of construction that depends on a political conception of persons as free and equal citizens. Individuals who endorse this conception recognize themselves as endowed with two moral powers: first, they have a capacity for a sense of justice; second, they have a capacity to rationally pursue a conception of the good.[15] This directly leads to the construction of the original position, where members of N^* choose principles of justice in ignorance of their personal positions and characteristics while endorsing the thin theory of the good. Letting RN be the set of all-things-considered reasons that can justify the endorsement of principles of justice in N, the permissible reasons are restricted to a subset $R^* \subset RN$. If P is the set of all possible conceptions of justice, the reasons R^* available to the members of N^* put in circumstances C^* make them select a subset P^* of *liberal conceptions of justice*. These conceptions share a small set of features.[16] This first justificatory stage is already public in the sense that the set R^* is public because it follows implicitly from the public recognition that persons are free and equal citizens. In the contrary case, the original position could not serve as a device of justification among the members of N. The justification of P^* is only a *pro tanto* justification, however.

To achieve *full justification*, we need to consider whether components of P^* are still justified when transitioning from the triple $\{N^*, C^*, R^*\}$ to the triple $\{N, C, RN\}$.[17] The issue is whether each citizen can individually and rationally endorse the conceptions contained in P^* in light of their comprehensive doctrine, especially their conception of the good. In the most permissive, "wide" version of public reason Rawls exposes, comprehensive doctrines can reasonably be introduced at the stage of full justification, provided that public reasons are presented in due course in case of conflict between comprehensive doctrines. More formally, full justification requires that if $\{N^*, C^*, R^*\} \Rightarrow P^*$,

characterization of the RMPR is consistent with the two-part version of their model III. Their one-part version of model III and their model IV stretch Rawls's political liberalism toward the diversity-convergence account of stability, probably too much.

15 Rawls, *Political Liberalism*, 18.

16 Rawls, "The Idea of Public Reason Revisited," 773. Crucially, though Rawls obviously sees justice as fairness as belonging to P^*, he leaves open the possibility that P^* is not a singleton.

17 "Full justification is carried out by an individual citizen as a member of civil society. . . . In this case, the citizen accepts a political conception and fills out its justification by embedding it in some way into the citizen's comprehensive doctrine as either true or reasonable, depending on what that doctrine allows" (Rawls, *Political Liberalism*, 386).

then $\{N, C, RN\} \Rightarrow P^*$, with the proviso that $\{N, C, R^*\} \Rightarrow P^*$. If this condition is satisfied, an overlapping consensus is established: the liberal conceptions of justice contained in P^* are all endorsable from within the citizens' comprehensive doctrines. That means that members of N's all-things-considered reasons RN sufficiently overlap to provide shared support for these conceptions. This solves the first part of the stability problem, the justificatory instability problem.

A last step is required to solve the second part of the stability problem, the assurance problem. Recall here that the problem is not whether persons have reasons to act rightly but whether they can have the assurance that others will do so. As I explain above, the assurance requirement trickles up to the whole hierarchy of belief: Ann must have sufficient reason to believe that Bob will act rightly, which requires that she have sufficient reason to believe that Bob has sufficient to reason to believe that Ann will act rightly, and so on. That implies that the assurance problem cannot be solved unless the existence of an overlapping consensus is itself a *public event* in N. By definition, a public event is commonly known.[18] In this case, it is commonly known among members of N that conceptions in P^* are fully justified for reasons R^* that are at the same time shared and endorsable from within the various comprehensive doctrines. The assurance problem is solved because the public establishment of the overlapping consensus (which solves the justificatory instability problem) makes it common knowledge that liberal conceptions of justice are endorsed for shared public reasons. Stability for the right reasons is achieved because officials and citizens all justifiably believe that everyone has shared sufficient reasons to act rightly.

3. The Diversity-Convergence Critique of the RMPR

The literature on public reason is huge, and the Rawlsian account of public reason has been the target of a long list of criticisms.[19] I shall not discuss all of them here. Rather, I will focus on the criticisms that explicitly and specifically challenge the ability of the RMPR to solve the stability problem. These criticisms share at least two features. First, they mostly ignore the justificatory instability problem—implicitly indicating that the RMPR provides a satisfactory answer to it. Instead, they argue that public reason cannot be the assurance device that the RMPR assumes it is. Second, they tend to favor a convergence account

18 For an insightful analysis of the role of public events in fostering common knowledge in a variety of social situations, see Chwe, *Rational Ritual*. For a formal account, see Milgrom, "An Axiomatic Characterization of Common Knowledge."

19 See Quong, *Liberalism without Perfection*, 259–60, for a survey of this list.

of public reason, instead of Rawls's consensus approach.[20] Proponents of this account tend to argue that Rawls's consensus approach depends on unrealistic or normatively doubtful assumptions to solve the stability problem in highly diverse societies. I present and explore in this section the various arguments of the diversity-convergence critique against the RMPR.

3.1. Public Reason Is Cheap Talk

A first argument against the Rawlsian solution to the stability problem is that public reasons cannot provide the required assurance because they are merely *cheap talk*. Consider a specific relation between two public officials (e.g., a judge and a legislator), between an official and a citizen, or between a set of officials and a set of citizens.[21] The RMPR suggests that the use of public reasons in debates over fundamental political questions will assure others that one intends to act rightly. It would signal one's intention to sincerely abide by the liberal principles of justice. As several critics note, the use of public reasons cannot, however, serve as a proper signal in the context of an assurance game as depicted in figure 1.[22] The reason is that each person has an interest in making the other believe that she will act rightly, independently of her actual intention. Indeed, if person 2 believes that person 1 will act rightly, then person 2 has (assuming that the justificatory instability problem has been solved) all-things-considered reasons to act rightly too. Crucially, this guarantees to person 1 either of her two most preferred outcomes. Moreover, using public reasons in this model is costless. That means that one has absolutely no reason not to *pretend* to act rightly by using public reasons. This information is common knowledge among the players. That means that the use of public reasons does not lead to trustworthy messages and cannot solve the assurance problem.

3.2. Public Reasoning Is Vulnerable to Noise and Its Amplification

Thrasher and Vallier highlight a second problem with the RMPR related to its vulnerability to noise and its amplification.[23] Experimental evidence indicates that cheap talk can sometimes slightly favor cooperation in social dilemmas.

20 For the consensus/convergence distinction, see D'Agostino, *Free Public Reason*.

21 Rawls indicates that the ideal of public reason applies to government officials and candidates for public office ("The Idea of Public Reason Revisited," 766–67). Relations between citizens *per se* are outside the scope of public reason. See, however, Quong, *Liberalism without Perfection*, 273–75, for an argument in favor of a broader scope of public reason.

22 Gaus, "A Tale of Two Sets"; Kogelmann and Stich, "When Public Reason Fails Us"; Thrasher and Vallier, "The Fragility of Consensus."

23 Thraser and Vallier, "The Fragility of Consensus."

This is especially the case in face-to-face settings.[24] However, most of the interactions in the well-ordered society would take place under more impersonal settings. In this case, even if a global norm of mutual assurance prevails, the equilibrium (and thus the consensus) is susceptible to being destabilized by occasional—and possibly involuntary—defections. This can be the case because random errors can be interpreted as an intentional unwillingness to act rightly, leading to further defections. The result is an informational cascade leading to a sudden switch from a situation where almost everyone acts rightly to a situation where everyone disregards the principles of justice.[25] This problem is even more acute under the wide view of public reason. The wide view indeed tolerates the introduction in the political forum of nonpublic reasons attached to comprehensive doctrines. This makes it even harder to discriminate between insincere uses and sincere uses of private reasons that one intends to back up with public reasons.

3.3. Public Reason Is Incomplete and Manipulable

A third criticism underlines the incompleteness of public reason and its related manipulability.[26] Incompleteness can have two origins. On the one hand, public reason may be *indeterminate* because it cannot provide an answer to some fundamental political questions—for example, What should the monetary policy of the European Central Bank be? The answer to this question is unlikely to be found within the realm of public reason, at least if one accepts Rawls's requirement that it not appeal to controversial arguments not accessible to all citizens.[27] That seems to imply that citizens and officials must appeal to their comprehensive doctrines and nonpublic reasons to adjudicate this kind of issue. In these circumstances, public reason can no longer be used as an assurance mechanism. On the other hand, incompleteness can result from *inconclusiveness*. In this case, public reason provides contradictory and nondominated justifications on political issues. The worry is not that citizens and officials may appeal to nonpublic reasons but rather that they might choose public reasons that are the most favorable to their private interests. Again, this

24 Wilson and Sell, "'Liar, Liar . . .'"

25 Game-theoretic evolutionary models moreover show that in games with similar structures to the assurance game, only the suboptimal equilibrium is stochastically stable. See, for instance, Young, *Individual Strategy and Social Structure*.

26 Kogelmann, "Public Reason's Chaos Theorem."

27 Note, however, that Rawls is explicitly requiring the completeness of the *political conceptions* underlying public reason ("The Idea of Public Reason Revisited," 777). But presumably a political conception does not fully cover topics such as monetary policy, though monetary policy has an obvious relevance to distributive justice.

presumably considerably weakens the ability of public reason to serve as an assurance mechanism. Inconclusiveness indeed makes public reason not strategy-proof, and thus vulnerable to manipulation, along lines similar to standard results about strategy-proofness in social choice theory.[28]

3.4. Public Reason Depends on a Common Knowledge Condition

A fourth and last criticism emphasizes the problems related to the fact that common knowledge is a precondition for Rawlsian stability for the right reason.[29] As I explain above, the transition from full justification to public justification implies that the existence of the overlapping consensus is common knowledge among the members of *N*. At least, this is required if public reason is to serve as an assurance mechanism. The assurance has to be given across the whole belief hierarchy of each official and citizen.[30] This requirement, however, makes public reason very fragile as an assurance mechanism as soon as there is a small fraction of unreasonable persons in *N*. The presence of a fraction ε of unreasonable persons in *N* means that reasonable persons are confronted with a Bayesian game: with a probability of $1 - \varepsilon$ they play the assurance game of figure 1, but with probability ε they are part of an assurance dilemma with an unreasonable person (fig. 2). Acting wrongly is the dominant strategy for the unreasonable person. It is then best for the reasonable person to also act wrongly.

		Person 2 (unreasonable)	
		Act rightly	Act wrongly
Person 1 (reasonable)	Act rightly	3; 2	0; 3
	Act wrongly	2; 0	1; 1

FIGURE 2 An Assurance Dilemma

We might assume that knowing the other's type is sufficient for a reasonable person to choose to act rightly (if the other's type is reasonable). But this is not the case: the reasonable person has to know that the other person knows that she is reasonable, and so on. Ultimately, public reason can serve as an assurance mechanism only if it is common knowledge that both persons are reasonable.

28 Taylor, *Social Choice and the Mathematics of Manipulation*.

29 Chung, "The Instability of John Rawls's 'Stability for the Right Reasons.'"

30 As Chung notes, this is never made explicit in Rawls's writings ("The Instability of John Rawls's 'Stability for the Right Reasons'"). Readers of Rawls who endorse his public reason model recognize, however, the common knowledge requirement. See Hadfield and Macedo, "Rational Reasonableness"; and Weithman, *Why Political Liberalism?*

As Chung formally demonstrates, whenever there is a fraction ε arbitrarily close to zero of unreasonable persons, there is always a Bayesian game where both persons acting wrongly is the unique equilibrium.[31]

4. THE RMPR AND COMMUNITY-BASED REASONING

We may ask whether the RMPR can be adapted to respond to the objections that have been leveled against it by the diversity-convergence account. The questions at stake are the following: Under what condition(s) can public reason serve as an assurance mechanism? Are these conditions likely to apply in diverse liberal societies?

Based on a coordination model that Hadfield and Weingast developed, Hadfield and Macedo argue that public reason can solve the assurance problem by fostering a common logic among the members of the relevant population.[32] This common logic provides the basis for believing that everyone will act rightly. In Hadfield and Weingast's model, a third-party institution supplies the common logic in a population of buyers and sellers who want to trade for goods. The third-party institution provides a classification of agents' behavior, in particular whether the action of an agent can be classified as "cheating." Based on this classification, the institution supports a system of belief such that agents are credibly threatened with being punished if they cheat. On this basis, buyers are able to coordinate on boycotting sellers who would cheat, thus deterring actual cheating behavior from the sellers. According to Hadfield and Macedo, this model "helps us understand [that] the importance of public reason is the recognition that in order to coordinate the participants in a community dependent on a decentralized enforcement of rules and principles expressing judgments of right and wrong, the common logic must be *publicly accessible*. Indeed, in game-theoretic terms, it must be *common knowledge*."[33] Hadfield and Macedo go on to emphasize that endorsing the publicly accessible common logic could ultimately lead persons to set aside their personal reasons and inference norms, and accept a form of normativity derived from a "we-mode" of thinking.[34]

So far, so good. Hadfield and Macedo's account appropriately captures the common knowledge requirement that I emphasize above. However, they do

31 Chung, "The Instability of John Rawls's 'Stability for the Right Reasons.'" This result can be seen as a variant of Rubinstein's "electronic mail game" ("The Electronic Mail Game"). This kind of model shows that common knowledge and approximation of common knowledge have very different implications.

32 Hadfield and Weingast, "What Is Law?"; Hadfield and Macedo, "Rational Reasonableness."

33 Hadfield and Macedo, "Rational Reasonableness," 12–13, emphasis in original.

34 Hadfield and Macedo, "Rational Reasonableness," 39.

not make it clear how public reason makes the common logic publicly available. Hadfield and Weingast's model *postulates* the existence of a third-party institution that makes public announcements. It does not give any indication regarding the circumstances under which this institution emerges and is able to achieve the required publicness. I shall now propose an extension of the RMPR that accounts for a plausible form of practical reasoning through which public reason generates common knowledge. Following my earlier work, I call this form of practical reasoning *community-based reasoning*.[35] It can be characterized in the following way:

> *Community-Based Reasoning*: A person *i*'s practical reasoning is community-based if, in some strategic interaction *S*, her action *A* follows from the following reasoning steps:
>
> 1. *i* believes that [she] and all other persons *j* in *S* are members of some community *C*.
> 2. *i* believes that some state of affairs *x* holds.
> 3. Given 2, 1 grounds *i*'s belief that all other persons *j* in *S* believe that *x* holds.
> 4. From *x*, *i* inductively infers that some state of affairs *y* also holds.
> 5. Given 3 and 4, 1 grounds *i*'s belief that all other persons *j* in *S* believe that *y* also holds.
> 6. From 5, and given *i*'s preferences, *i* concludes that *A* is best in *S*.[36]

Suitably reformulated in the terms of the RMPR, we can say that members of *N* are community-based reasoners if they infer that they have all-things-considered reasons to act rightly partly from the fact that they belong to the same *political community*. Community membership indeed sustains two key steps in this logic of reasoning. In step 3, community membership serves as a basis for *i* to infer that *x* is mutual knowledge. In step 5, it serves as a basis for *i* to infer that everyone in *N* infers *y* from *x*. The term "grounds" in both steps singles out the fact that *i* assumes, based on community membership, that she and other members of *N* are sharing both privileged epistemic accessibility to some state of affairs *x* and some form of inductive inference from *x* to *y*.

The key point here is that if all members of *N* are community-based reasoners with respect to some state of affairs *x*, then it can be shown that *y* is common

35 Hédoin, "A Framework for Community-Based Salience" and "Community-Based Reasoning in Games." I leave most of the technical details aside. The interested reader may consult "Community-Based Reasoning in Games," in which an epistemic game-theoretic model is presented.

36 Hédoin, "Community-Based Reasoning in Games," 4.

belief (or knowledge) in *N*.[37] Substitute "everyone else acts rightly" for *y*; *x* refers to a mutually accessible state of affairs or event—for example, an official's making an announcement in an assembly. Now, upon observing *x*, any member of *N* will first infer that everyone else, as members of the same community, has also observed *x*. She will also infer from *x* that *y* is the case, and that everyone else makes the same inference from *x* to *y*. Then, she believes that *y* is mutual belief in *N*. Therefore, one can infer from *x* that *y* is mutual belief in *N*, which we denote by *y′*. Substitute *y′* for *y* in step 4 above. From 5, one can infer that *y′* is mutual belief in *N*, which we denote *y″*. We can reiterate the process indefinitely, thus establishing that *y* is common knowledge in *N*. Assuming that the members of *N* are community-based reasoners then solves the intricating issue of the origins of common knowledge. It thus directly answers the objection discussed in section 3.4 above.

Provided that the justificatory instability problem has already been resolved, community-based reasoning also responds to the cheap talk objection. Once person *i* has reached the conclusion that it is common knowledge that everyone will act rightly, the preference structure of the assurance game makes acting rightly the strict best response (see fig. 1). Suppose that *x* denotes some announcement, based on one or several public reasons, made by a public official in an assembly. Two points are relevant here. First, we implicitly have to assume that everyone agrees on what constitutes "public reasons." This is what lurks behind step 3. This assumption is directly grounded on the fact that because members of *N* are members of the same community, we consider that *by definition* they share a common conception of what public reasons are. Second, once again by the very fact that they are members of the same community, we assume that members of *N* inductively infer from *x* that everyone else will act rightly. Strictly speaking, this inference may be false. It could happen that some persons in *N* do not act justly even if they observe *x*. But in this very community, that will not happen. Members of *N* just share a *lebensform*, a "form of life" in the Wittgensteinian sense. This is sufficient to provide the required assurance. In this community, using public reasons is not cheap talk, as a matter of fact. Quite the contrary, the use of public reasons is a *community-based salient event* in the sense I offer elsewhere: an event based on which everyone correctly infers that some proposition is commonly known among a community.[38] This very

37 See Hédoin, "Community-Based Reasoning in Games," 7–8, for a formal proof. Lewis is the first to have proposed an account that shows how a proposition *becomes* common knowledge in a population of agents who share a form of inductive inference and have mutual epistemic access to a given state of affairs (*Convention*).

38 Hédoin, "Community-Based Reasoning in Games," 10.

property singles out this event, separating it from the stream of events that any normally rational person can observe.

Community-based reasoning thus provides a sufficient condition for public reason to satisfy the required common knowledge condition. This account adequately supplements Hadfield and Macedo's claim that public reason makes common knowledge a common logic in a population.[39] Two remarks should be made in addition. The first remark is related to a hidden assumption behind the above argument: here, I have implicitly assumed not only that everyone in N is a community-based reasoner but also that it is somehow common knowledge that this is the case. This is indeed required to make the iterative steps that successively show that y, y', y'', and so on, are mutual knowledge. Where does this common knowledge come from? A plausible answer, at least in some circumstances, is just to point out that community membership is a public event *and* that to be a member of a community is to share some form of reasoning, including inductive inference, through common practices. In this case, community-based reasoning can indeed become common knowledge in the relevant population.[40] This leads to an ostensible Wittgensteinian reinterpretation of the concept of public reason.[41] To make use of public reasons is to follow some rules that, at the bottom, correspond to shared practices that do not have any fundamental justification. Public reason is just what members of a community agree it is through their practices, and to share a conception of public reason and the related practices is just what it is to be a member of some community. This emphasizes an important link between justification and stability. To be fully justified, principles of justice P^* must be agreed on by members of the relevant population, based on public reasons R^*. One is then a member of the community by virtue of agreeing on P^* based on R^*. Agreeing on the justification of P^* based on R^* defines at the same time what it is to be a member of the community. In this Wittgensteinian sense, if principles P^* are properly justified, they must already be common knowledge. The step from full to public justification is merely theoretical; in practice, they are one and the same thing.

39 Hadfield and Macedo, "Rational Reasonableness."

40 See note 18 above.

41 See especially Wittgenstein's remarks on "language games" and "forms of life," pointing out the relationship between meaning and the existence of shared practices among the members of a community (*Philosophical Investigations*). As Forrester has documented, Rawls's early social philosophy has been largely influenced by Wittgenstein's conception of rules and games (*In the Shadow of Justice*). It should be acknowledged, however, that this reinterpretation is only partial. A full Wittgensteinian account of public reason would assert that *all* reasons are public.

The second point is more exegetical. Rawls explicitly states that a "well-ordered democratic society is *neither a community* nor, more generally, an association."[42] Rawls notices two differences between a well-ordered society and an association. On the one hand, a well-ordered society is closed in the sense that one cannot enter or exit it in the course of her life. On the other hand, a well-ordered society does not have final ends or aims. Moreover, a well-ordered society "is not a community either, if we mean by a community a society governed by a shared comprehensive religious, philosophical, or moral doctrine." He adds that this "fact is crucial for a well-ordered society's idea of public reason. To think of a democracy as a community (so defined) overlooks the limited scope of its public reason founded on a political conception of justice."[43] These remarks can be related to the distinction between the political and the comprehensive that is at the core of political liberalism, and to the exclusion of comprehensive reasons from the realm of public reason. The point is that to solve the stability problem, there must be *something* common and public in the relevant population. Rawls may have been wrong in excluding truth, perfectionist values, and more generally comprehensive reasons from the realm of public reason. On the other hand, from a Rawlsian perspective, the burdens of judgment make unlikely the possibility that reasons based on metaphysical, religious, or moral considerations can serve as a basis for public justification. The difficulty, pointed out multiple times in the literature, is that it is unclear why the disagreement between persons stemming from the burdens of judgment should not concern political principles of justice. Rawls's answer to this objection is to argue that the idea of public reason is constitutive of a constitutional democratic society governed according to liberal principles of justice. The idea of public reason has then presumably no bearing in societies that are not *de facto* liberal and democratic in this sense. If we follow Rawls in this postulate, we may then argue that the members of N belong to a *democratic political* community. This political community is defined by the fact that its members accept a specific political conception of justice, along the lines Rawls identified.[44]

5. Some Concerns and Objections

I shall address in this section some concerns and objections to my proposed interpretation of the RMPR in terms of community-based reasoning. This will also permit making more explicit its implications and filling in some detail.

42 Rawls, *Political Liberalism*, 40, emphasis added.

43 Rawls, *Political Liberalism*, 41.

44 Rawls uses the term "political community" in *Political Liberalism*, but not in this sense.

There are, in particular, two related worries: that reinterpreting the RMPR in terms of community-based reasoning makes too strong a concession to communitarianism and that it depends on a too-demanding form of normalization.[45] These worries find their roots in the above characterization of community-based reasoning. In particular, it might be suggested that steps 3 and 5 make strong demands on the form of practical reasoning required and that, because of that, it will either fail to solve the stability problem most of the time or do so in a manner that betrays the very point of Rawls's political liberalism. Another way to state this concern is to ask what remains of the idea of public reason as a distinctive account to solve the stability problem under the community-based approach. I shall answer in two steps.

There is no doubt that the proposed revision of the RMPR goes in the direction of communitarianism. This is not only because the account of community-based reasoning obviously relies on the concept of community. This is more fundamentally related to the fact that the idea of public reason, at least in Rawls's political liberalism, is tightly related to legitimacy and justification.[46] I have acknowledged above that under this revised interpretation of the RMPR, using public reason consists in following rules that do not have a fundamental justification, except that they are constitutive of shared social practices. Now, if the community-based reasoning account indeed entails deflationary concepts of justification and legitimacy, this is not *totally* inconsistent with Rawls's political turn. As I have noted in the preceding section, that turn itself strongly suggests that the idea of public reason is grounded in what can be called the *democratic form of life*, consisting of political and social practices that are themselves a reflection of the *democratic political culture*. In other words, there is already in Rawls's late solution to the stability problem the admission that public reason, and more generally the whole political conception, is related to some way of life that is specific to liberal democratic societies. My argument for the community-based reasoning reinterpretation of the RMPR can then be formulated as follows: as a device that makes it possible for a society to foster stability around a conception of justice, public reason is specific to the democratic form of life of liberal democracies but still has to rely on a form of practical reasoning that is community based—a form of practical reasoning that is not specific to any kind of society but that we may well see at work in any human community. This form of practical reasoning makes it possible to satisfy the requirement of common knowledge that, as we have seen, is needed for public reason to fulfill its job.

45 I thank the two anonymous reviewers who have each and in different ways pushed me to be more explicit about these worries and how they can be addressed.

46 Rawls's "liberal principle of legitimacy" is indeed stated and defended in the context of his account of public reason (*Political Liberalism*, 216).

This leads to the second worry. Is it plausible to think that this revised version of the RMPR can account for the way public reason solves the stability problem—if indeed it does—in modern liberal democracies? I should first mention that community-based reasoning is only a partial solution to the stability problem. It depends on the fact that persons indeed are community-based reasoners. The plausibility of this assumption depends not only on empirical considerations but also on what we regard as the appropriate analytical understanding of the concept of community. The risk is to navigate between the tautology (and thus empirical irrelevance) and the empirically refuted. Regarding the latter, it may indeed happen to be that contemporary societies are too diverse and, as a consequence, that community-based reasoning is not available as a form of practical reasoning. It can be argued in particular that it is unlikely that the members of a nation form a community in the relevant sense. This is of course a relevant and fully justified worry. There is no doubt that in some polities the lack of shared national identity can be a major obstacle to the emergence of a stable overlapping consensus around the constitutional essentials and matters of basic justice that are the subject of Rawlsian public reason.[47] Even in more culturally homogeneous polities, it is unclear that practical reasoning can be community based. We should, however, remember that, at least in Rawls's version, the idea of public reason is limited in scope and content. It concerns only the "political" aspect of constitutional essentials and matters of basic justice, and it applies only to discussions within the "public political forum": the discourse of judges in their decisions, the discourse of government officials, and the discourse of candidates for public office.[48] Within this restricted scope, the assumption that persons—for instance, members of parliament—are at least sometimes community-based reasoners seems to be less demanding.[49]

Even if this answers the two major worries I have identified, it should be acknowledged again that community-based reasoning remains only a partial solution to the stability problem. In the preceding section, I argue that community-based reasoning allows the RMPR to solve two objections: that public

47 In a worse case, this lack can escalate into a civil war. But less dramatic examples, such as the case of contemporary Belgium, illustrate the kind of difficulties that may arise and how they can be interpreted from the Rawlsian perspective.

48 Rawls, "The Idea of Public Reason Revisited," 767.

49 Rawls notes that citizens, when they vote on constitutional essentials and matters of basic justice, "are to think of themselves *as if* they were legislators" and so must use the requirements of public reason to assess government officials and candidates for public office ("The Idea of Public Reason Revisited," 769). We may doubt that this corresponds to actual practice in liberal democracies. Indeed, this demand is more akin to what Rawls calls the *ideal* than the idea of public reason.

reason is cheap talk and that it depends on a common knowledge condition. I have said nothing, however, about the other two objections: that public reason is vulnerable to noise and to manipulation. Moreover, as Chung shows, satisfying the common knowledge condition is far less easy if the justificatory instability problem has not yet been fully solved.[50]

This points to a simple but essential fact: stability with too much diversity is compromised. In other words, there must be something common among the members of a society to obtain the required stability—that is, not a mere *modus vivendi*. Ideally, this common minimal basis must be furnished by the political institutions themselves—through education, for instance—as Rawls himself underlined. More generally, by serving the role of correlating devices, social norms governing relationships between public officials, and between public officials and citizens, provide the basis on which public reason operates in a well-ordered society. Particular norms or rules do not select directly a particular political conception. But as consistent systems of norms and rules, institutions create *indirect public reasons* to settle on a particular conception in specific circumstances.[51] But the existence of these norms is itself constitutive of forms of life—that is shared modes of reasoning generating focal points and salient events. As Thrasher and Vallier explicitly admit, their model "assumes that the relevant type of stability [i.e., stability in the Rawlsian sense] is already in place."[52] This is a severe limit because as a result it hardly improves on the RMPR.[53] Arguably, the more a society is diverse in its practices and beliefs, the smaller the scope of public reason will be. At the extreme, public reason must rely on the lowest common denominator that makes people belong to the same society.

6. CONCLUSION

Proponents of the diversity-convergence approach do not stop at criticizing Rawls's model of public reason. They have also argued that a convergence account of public reason where persons agree over rules and principles *for*

50 Chung, "The Instability of John Rawls's 'Stability for the Right Reasons.'"

51 Thrasher and Vallier, "The Fragility of Consensus."

52 Thrasher and Vallier, "The Fragility of Consensus," 950.

53 As they seem to implicitly admit, Thrasher and Vallier's model is not incompatible with the RMPR ("The Fragility of Consensus"). There is no obvious reason why the latter would not admit the role played by "choreographers" in fostering social coordination. Thrasher and Vallier underestimate the importance of shared reasoning for the existence of correlated equilibria, which is formally captured by the so-called common prior assumption (see Gintis, *The Bounds of Reason*, ch. 7).

different reasons solves the stability problem.[54] It is beyond the scope of this paper to assess the merits of the convergence account in comparison with those of the RMPR with respect to the stability problem. To end this essay, however, two brief remarks are in order. First, what I have just said above suggests that a pure convergence account is unlikely to succeed, at least if more than a mere *modus vivendi* is aimed at. As Chung points out using a formal argument, to the extent that diversity promotes extremist views, it is not the case that diversity will strengthen stability, contrary to what, for instance, Kogelmann and Stich argue.[55] Second, it is not clear what remains "public" in an account of public reason where persons are permitted to bring any beliefs and reasons in support of a law. This is not a mere terminological quibble; it also questions the very nature of the social order in liberal democracies. Overall, this is the whole project of public reason liberalism that must be reconsidered in light of the difficulty in solving the stability problem.[56]

University of Reims Champagne-Ardenne
cyril.hedoin@univ-reims.fr

REFERENCES

Binmore, K. G. *Just Playing*. Vol. 2 of *Game Theory and the Social Contract*. Cambridge, MA: MIT Press, 1998.

Broome, John. *Weighing Goods: Equality, Uncertainty and Time*. Oxford: Blackwell, 1991.

Chung, Hun. "The Instability of John Rawls's 'Stability for the Right Reasons.'" *Episteme* 16, no. 1 (March 2019): 1–17.

———. "The Well-Ordered Society under Crisis: A Formal Analysis of Public Reason vs. Convergence Discourse." *American Journal of Political Science* 64, no. 1 (January 2020): 82–101.

Chwe, Michael Suk-Young. *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton, NJ: Princeton University Press, 2003.

54 Gaus, *The Order of Public Reason*; Kogelmann and Stich, "When Public Reason Fails Us."

55 Chung, "The Well-Ordered Society under Crisis"; Kogelmann and Stich, "When Public Reason Fails Us."

56 This essay was presented at the STOREP conference in June 2021 (held online) and at the international conference "Héritages et usages de Rawls: *Théorie de la justice*, 50 ans après," held in November 2021 in Paris. I received valuable comments from participants at these events. I thank especially Samuel Ferey and Alain Marciano. I also thank two anonymous reviewers for their remarks.

D'Agostino, Fred. *Free Public Reason: Making It Up as We Go*. New York: Oxford University Press, 1996.

Forrester, Katrina. *In the Shadow of Justice: Postwar Liberalism and the Remaking of Political Philosophy*. Princeton, NJ: Princeton University Press, 2019.

Gaus, Gerald F. *Justificatory Liberalism: An Essay on Epistemology and Political Theory*. New York: Oxford University Press, 1996.

———. "Moral Conflict and Prudential Agreement: Michael Moehler's *Minimal Morality*." *Analysis* 79, no. 1 (January 2019): 106–15.

———. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. New York: Cambridge University Press, 2011.

———. "A Tale of Two Sets: Public Reason in Equilibrium." *Public Affairs Quarterly* 25, no. 4 (October 2011): 305–25.

Gaus, Gerald, and Chad Van Schoelandt. "Consensus on What? Convergence for What? Four Models of Political Liberalism." *Ethics* 128, no. 1 (October 2017): 145–72.

Gintis, Herbert. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton, NJ: Princeton University Press, 2009.

Granovetter, Mark. "Threshold Models of Collective Behavior." *American Journal of Sociology* 83, no. 6 (May 1978): 1420–43.

Hadfield, Gillian K., and Stephen Macedo. "Rational Reasonableness: Toward a Positive Theory of Public Reason." *Law and Ethics of Human Rights* 6, no. 1 (2012): 7–46.

Hadfield, Gillian K., and Barry R. Weingast. "What Is Law? A Coordination Model of the Characteristics of Legal Order." *Journal of Legal Analysis* 4, no. 2 (Winter 2012): 471–514.

Hédoin, Cyril. "Community-Based Reasoning in Games: Salience, Rule-Following, and Counterfactuals." *Games* 7, no. 4 (December 2016): 36.

———. "A Framework for Community-Based Salience: Common Knowledge, Common Understanding and Community Membership." *Economics and Philosophy* 30, no. 3 (November 2014): 365–95.

———. "Naturalism and Moral Conventionalism: A Critical Appraisal of Binmore's Account of Fairness." *Erasmus Journal for Philosophy and Economics* 11, no. 1 (Spring 2018): 50–79.

Kogelmann, Brian. "Public Reason's Chaos Theorem." *Episteme* 16, no. 2 (June 2019): 200–219.

Kogelmann, Brian, and Stephen G. W. Stich. "When Public Reason Fails Us: Convergence Discourse as Blood Oath." *American Political Science Review* 110, no. 4 (November 2016): 717–30.

Lewis, David K. *Convention: A Philosophical Study*. Oxford: Blackwell, 2002.

Milgrom, Paul. "An Axiomatic Characterization of Common Knowledge."

Econometrica 49, no. 1 (January 1981): 219–22.
Quong, Jonathan. *Liberalism without Perfection*. Oxford: Oxford University Press, 2011.
Rawls, John. "The Domain of the Political and Overlapping Consensus." *New York University Law Review* 64, no. 2 (1989): 233–55.
———. "The Idea of Public Reason Revisited." *University of Chicago Law Review* 64, no. 3 (Summer 1997): 765–807.
———. "Justice as Fairness: Political Not Metaphysical." *Philosophy and Public Affairs* 14, no. 3 (Summer 1985): 223–51.
———. "Kantian Constructivism in Moral Theory." *Journal of Philosophy* 77, no. 9 (September 1980): 515–72.
———. *Political Liberalism*. New York: Columbia University Press, 1993.
———. *A Theory of Justice*. Cambridge, MA: Belknap Press, 1971.
Rubinstein, Ariel. "The Electronic Mail Game: Strategic Behavior Under 'Almost Common Knowledge.'" *American Economic Review* 79, no. 3 (June 1989): 385–91.
Sugden, Robert. *The Economics of Rights, Co-operation and Welfare*. 2nd ed. New York: Palgrave Macmillan, 2005.
Taylor, Alan D. *Social Choice and the Mathematics of Manipulation*. Cambridge: Cambridge University Press, 2005.
Thrasher, John. "Agreeing to Disagree: Diversity, Political Contractualism, and the Open Society." *Journal of Politics* 82, no. 3 (July 2020): 1142–55.
Thrasher, John, and Kevin Vallier. "The Fragility of Consensus: Public Reason, Diversity and Stability." *European Journal of Philosophy* 23, no. 4 (December 2015): 933–54.
———. "Political Stability in the Open Society." *American Journal of Political Science* 62, no. 2 (April 2018): 398–409.
Vanderschraaf, Peter. *Strategic Justice: Convention and Problems of Balancing Divergent Interests*. Oxford: Oxford University Press, 2019.
Weithman, Paul. *Why Political Liberalism? On John Rawls's Political Turn*. New York: Oxford University Press, 2011.
Wilson, Rick K., and Jane Sell. "'Liar, Liar . . .': Cheap Talk and Reputation in Repeated Public Goods Settings." *Journal of Conflict Resolution* 41, no. 5 (October 1997): 695–717.
Wittgenstein, Ludwig. *Philosophical Investigations*. Malden, MA: Wiley-Blackwell, 2010.
Young, H. Peyton. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press, 2001.