AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

The Well-Ordered Society under Crisis: A Formal Analysis of Public Reason vs. Convergence Discourse

Hun Chung Waseda University

Abstract: *A well-ordered society faces a crisis whenever a sufficient number of noncompliers enter into the political system. This has the potential to destabilize liberal democratic political order. This article provides a formal analysis of two competing solutions to the problem of political stability offered in the public reason liberalism literature—namely, using public reason or using convergence discourse to restore liberal democratic political order in the well-ordered society. The formal analyses offered in this article show that using public reason fails completely, and using convergent discourse, although doing better, has its own critical limitations that have not been previously recognized properly.*

A problem that confronts us whenever we try to establish a social order is dealing with noncompliers. John Rawls deliberately tried to avoid this issue by restricting his theoretical task to providing principles of justice exclusively for a well-ordered society in which everybody "strictly complies with, and so abides by, the principles of justice" (Rawls 2001, 13). To this, Rawls makes it clear that his focus was on "ideal theory" rather than "non-ideal theory" (ibid.).

The fact that Rawls focused on "ideal theory" does not mean that he did not take the issue of real-world stability seriously. His turn to political liberalism was primarily motivated to show how "there may exist over time a stable and just society of free and equal citizens profoundly divided by reasonable religious, philosophical, and moral doctrines" (Rawls 1993/2005, xxv). His solution was to reinterpret justice as fairness as a freestanding, political conception of justice, which may first be endorsed by public reason alone—that is, by reasons that everybody considered only as free and equal democratic citizens can all endorse—and then be further supported by an overlapping consensus based on each citizen's comprehensive (yet reasonable) religious, philosophical, and moral doctrines.

The type of political stability thus achieved is what Rawls called "stability for the right reasons" (Rawls 1993/2005, 458–60). Stability for the right reasons is achieved when everybody endorses and is morally motivated, by her sense of justice, to follow the requirements of the political conception of justice justified from the same public reasons shared by all seen as free and equal democratic citizens and, furthermore, when this fact is known to all as a matter of common knowledge. We can see that "stability for the right reasons" is a highly idealized conception of political stability; it is in sharp contrast with a somewhat more realistic conception of political stability, which Rawls called *modus vivendi*—in which political stability is achieved by a (mere) balance of powers.

Rawls does have an explanation of how a just political institution that is stable for the right reasons, once established, can be properly sustained over time. According to Rawls, "those who grow up under just basic institutions—institutions that justice as fairness itself enjoins—acquire a reasoned and informed allegiance to those institutions sufficient to render them stable" (Rawls 2001, 185). However, one should note that the educational effects of a well-ordered society cannot be a complete solution to achieving stability for the right reasons. This is not because (as many have argued) the well-ordered society faces the problem of assurance even among reasonable people, whose motivation to cooperate

Hun Chung is Associate Professor, Faculty of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo, Japan 169-8050.

I thank Adam Gjesdal, Alex Shaefer, Brian Kogelmann, David Wiens, Jerry Gaus, Kevin Vallier, Sean Ingham, and Stephen Stich for providing helpful feedback and correcting typos on earlier drafts of this paper. I also thank the participants of the 'game theory group' organized by the PhD students in the Department of Philosophy at the University of Arizona for reading, presenting, and discussing this paper before it got published. All remaining faults are my own.

American Journal of Political Science, Vol. 00, No. 0, xxxx 2019, Pp. 1–20

DOI: 10.1111/ajps.12445

is conditional upon other people's cooperation.[1] The reason why Rawls's educational solution is incomplete is because it does not consider how it could cope with the intrusion of noncooperators, who may not have been brought up in a well-ordered society to benefit from its educational effects.[2] Such an intrusion can always happen through a random exogenous shock—such as when a wave of outside refugees enter into society by some unexpected political event—or by internal mutation. What we want from a theory of political stability is for it to show how a just society can be resilient against political turmoil caused by the intrusion of such noncompliers. We want a theory of political stability that shows how our well-ordered society can *restore* social *justice once it is destabilized*. If a well-ordered society lacks this sort of resilience, then the political stability that was initially achieved in a well-ordered society would be too fragile to serve as a foundation for modern liberal democratic institutions.

Let us say that a well-ordered society faces a *crisis* whenever a non-negligible proportion of rational but unreasonable people sufficient to destabilize liberal democratic political order have entered into the main decision-making bodies of the political system. The question that needs to be addressed, then, is how a well-ordered society under crisis can restore social justice once destabilized.[3]

There seem to be two main solutions that have been offered in the public reason liberalism literature: (a) using *public reason* and (b) using *convergence discourse*. According to the proponents of public reason, the way public reason can help maintain political stability in a liberal democracy is by serving as a reliable communicative signal that assures other reasonable political officials to cooperate their way toward social justice. That is, by only offering shared public reasons when discussing political matters, a public official is, in effect, sending a signal to other public officials that she is a reasonable person who fully endorses the shared political conception of justice of their liberal democratic society. The thought is that based on such public signals, other reasonable political officials will be assured that they are interacting with other reasonable people like them and will thereby be able to restore and retain social justice by mutual political cooperation. For instance, Hadfield and Macedo (2012) argue that public reason can perform a role similar to that of a common logic (i.e., a publicly accessible standard of right and wrong that everybody, despite her own idiosyncratic private logic, shares) in the Hadfield-Weingast positive model[4] that helps people achieve a long-lasting stable order by means of decentralized enforcement.

Many critics have argued that public reason is very unlikely to perform such a stabilizing role, as it is merely "cheap talk"—"a costless or very inexpensive, nonbinding communication" (Thrasher and Vallier 2015, 11)—that rational but unreasonable noncompliers can readily copy to falsely assure other reasonable compliers to cooperate, which they can then exploit for the sake of advancing their own private agendas (Gaus 2011; Thrasher and Vallier 2015).

As a remedy, Gaus proposes what he calls the "convergence" (as opposed to "consensus") framework of public justification (in short, "convergence discourse"), in which the basic requirement for individuals to uphold the same political conception of justice based on shared public reason (i.e. the requirement for "stability for the right reasons") is relaxed. Instead, political stability is achieved as long as citizens each find sufficient reasons to support the political conception of justice from their own particular set of reasons, which may very well be based on their comprehensive moral, religious doctrines[5] (see Gaus 2011, section 2.5).[6]

According to Kogelmann and Stich (2016; K&S hereafter), such a convergence framework of public justification can provide the requisite institutional tools to solve the problem of political stability in the well-ordered society. The specific way convergence discourse achieves political stability is by being *costly*. This might at first sound a bit paradoxical. As explained, convergence discourse relaxes the requirement for what counts as supporting the

[1] I personally think that the problem of assurance among reasonable people (i.e., conditional cooperators) in a well-ordered society has been misemphasized. If there are only conditional cooperators in a well-ordered society, the assurance problem can actually be solved rather trivially by any player making the first move to cooperate. Then, the only optimal action left for the other player is to cooperate as well, and, hence, the unique subgame-perfect equilibrium of a well-ordered society is mutual cooperation.

[2] As a matter of fact, it is possible for Rawls's educational solution to have a negative impact on restoring political stability when there are noncompliers in society. This will become more apparent in a later section.

[3] To be clear, Rawls assumed his well-ordered society to be a "closed" society and sidestepped the issues related to immigration. The proportion of unreasonable people added to the population here need not be immigrants. They may be those who became unreasonable through internal mutation.

[4] See Hadfield and Weingast (2012).

[5] Of course, even in a convergence conception of political stability, Gaus claims that punishment will always be needed because of uncertainty, a certain number of defectors, trembling hands, and so on (Gaus 2011, section 3.2.2).

[6] In a different vein, Muldoon (2016) has recently argued that using public reason to achieve and maintain stability will not be successful, as public reason will not be able to accommodate diversity in people's different perspectives. To understand how perspectival diversity may affect the proper operation of public reason and the adjudicative role of liberal rights, see Chung & Kogelmann (2018) and Chung (2019).

same political conception of justice; instead of having to support the society's political conception of justice on the basis of the same set of shared public reasons, individuals may support such a conception from whatever reasons that stem from their own comprehensive doctrines. If so, would this not render public discourse based on convergence discourse *cheaper*?

What makes convergence discourse costly is not in its requirement to participate in public discourse, but rather *in convincing other people* to think that one is truly a supporter of political justice. The thought is that, within a convergence framework of public justification, to truly convince others to believe that one truly endorses the political conception of justice of one's society, one would have to exert a considerable amount of effort *to learn other people's comprehensive doctrines* to provide reasons that can appeal to other people, not merely *qua* free and equal democratic citizens, but also *qua* endorsers of a particular comprehensive doctrine (K&S 2016, 724–27). Being able to provide genuine reasons that could truly appeal to other people's comprehensive doctrine can clearly separate the genuinely reasonable cooperators from the mere posers; in K&S's own words, doing so serves as a "blood oath." In short, according to K&S, political stability can be restored and reachieved by using convergent discourse as *costly signals*.

Like many articles written in the public reason liberalism literature, K&S's analysis on convergence discourse is informed by modern game theory. Yet, the game-theoretic model they present in their article is a bit too simple to allow us to both properly understand and critically evaluate the potential role convergence discourse may play in achieving political stability in the well-ordered society. The model that K&S presents in their article is a simple assurance game (K&S 2016, 724, Figure 4), which involves neither private information nor a distinction of types. K&S explain that their modeling choice was based on considerations of simplicity.[7] However, if a game has only a single type of player, there is no room to apply such concepts as cheap talk or costly signals. This means that, technically speaking, K&S did not provide an appropriate game-theoretical model that matches their philosophical intuitions.

The general purpose of this article is to provide a formal analysis of the two competing solutions (viz. public reason vs. convergence discourse) to the problem of instability in Rawls's well-ordered society. The more specific purpose of this article is to provide a more accurate game-theoretic model that better represents the logic of convergence discourse, and to examine whether K&S's main philosophical claims can be preserved. In this sense, this article can be read as a critical reply to Kogelmann and Stich (2016).

The analysis that proceeds will be game-theoretic. The next section will provide a baseline model of a well-ordered society that faces a crisis, which includes both a reasonable and an unreasonable type. It will then endogenously determine the proportion of unreasonable noncompliers that is sufficient to destabilize liberal democratic political order. We will see that the proportion of unreasonable noncompliers needed for this to happen is below majority. In the subsequent section, we will investigate whether the well-ordered society under crisis can restore social justice and reachieve liberal democratic stability if we allow our public officials to communicate via public reason modeled as cheap talk. The model will confirm that the criticisms raised by Thrasher and Vallier (2015) and Gaus (2011)—namely, that public reason cannot reliably solve the problem of mutual assurance and political stability—are accurate. We then extend our baseline model to explore whether the problem can be solved, as K&S claim, by allowing our public officials to use convergence discourse as costly signals. The results of the formal model partially support K&S's philosophical conclusions. But they also give us many reasons not to be too overly optimistic.

Baseline Model of the Well-Ordered Society under Crisis

This section presents the baseline model of Rawls's well-ordered society facing a crisis. Suppose that by some unexpected political event, a surge of refugees has entered into our well-ordered society. Among them are people who wish to change the well-ordered society into a perfectionist state in accordance with their own particular comprehensive doctrine. Suppose that these people have garnered a non-negligible amount of political support and have successfully sent their representatives into the major decision-making bodies of their political system. As a result, our well-ordered society is now going through a constitutional change—a change that is practically irrevocable once made.

Let $N = \{1, 2\}$ be the set of two public officials whose mutual agreement would be sufficient to amend the constitution of our well-ordered society under crisis in a specific way. Let $T = \{r,\ u\}$ be the set of types for public official 1, where r stands for "reasonable" and u stands for "unreasonable" in the Rawlsian sense. That is, if a

[7] K&S write: "For simplicity we assume that we are not working with non-polymorphic societies when it comes to preferences over acting on P and not acting on P" (2016, 718), meaning that their model has only one type of player.

public official is a reasonable type (i.e., type r), she has the requisite senses of justice and would be happy to live cooperatively under the fair terms of social cooperation proposed by the freestanding political conception of justice without having the desire to impose her particular comprehensive doctrine on other people. By contrast, if a public official is an unreasonable type (i.e., type u), she has a strong desire to coerce everybody to live under her particular comprehensive religious/moral doctrine that she believes to be correct. We assume that both the reasonable and unreasonable types are at least "rational" in the Rawlsian sense; that is, they have the capacity to both understand their own self-interests derived from their own particular conception of the good and choose the most effective means to achieve it (Rawls 1993/2005, 50).

To simplify the model, we assume that public official 2 is fixed as a reasonable type (i.e., type r.) However, we assume that public official 1 can either be reasonable or unreasonable (i.e., either type r or type u). So we are assuming one-sided incomplete information in which public official 1 has private information concerning her type that public official 2 does not fully know.

Each public official has two available actions: Cooperate (denoted by c) and Defect (denoted by d). Let $A_1 = A_2 = A = \{c,\ d\}$ be the set of actions for each political official $i = 1,\ 2$. When a political official cooperates, this means two things: (a) she proposes to amend the constitution in a way that fully honors the requirements of the political conception of justice regulating her society, and (b) she approves whatever constitutional amendment the other political official proposes. Conversely, when a political official defects, this means that (a) she proposes to amend the constitution with the specific intent to impose her particular comprehensive doctrine throughout society, and (b) she disapproves of whatever constitutional amendment the other political official proposes. I will not separately model the proposal and approval stages; proposing a reasonable constitutional amendment and approving the constitutional amendment proposed by the other political official are lumped together under the action label "cooperate," and proposing an unreasonable constitutional change and disapproving of the constitutional change proposed by the other political official are lumped together under the action label "defect." The main reason for doing so is because our current task is not to analyze the specific logistics of the constitution-making process, but to understand the destabilizing effect of the intrusions of noncompliers and whether this destabilization can be overcome by either public reason or convergence discourse.

A pure strategy for public official 1 is $s_1 : T \to A_1$, and a pure strategy for public official 2 is $s_2 \in A_2$. That is, a pure strategy for public official 1 is a function $s_1(t)$ that assigns an action in A_1 for each type $t \in T$, and a pure strategy for public official 2 is simply an action in A_2. (As a notational convention, I will sometimes write this as s_{1t}.) Let S_1 denote the set of all pure strategies of public official 1. Let ΔA_i denote the set of all probability distributions on A_i. Then a mixed strategy for public official 1 is $\sigma_1 : T \to \Delta A_1$ (as a notational convention, I will sometimes write this as σ_{1t}), and a mixed strategy for public official 2 is $\sigma_2 \in \Delta A_2$.

Let $O = \{$Hegemony, Justice, Instability, Subjugation$\}$ be the set of possible political outcomes in our well-ordered society. For each combination of actions performed by each political official, the following outcomes are generated:

- *Hegemony* when the political official herself unilaterally defects, whereas the other cooperates;
- *Justice* when both political officials cooperate;
- *Instability* when both political officials defect; and
- *Subjugation* when the political official herself unilaterally cooperates, whereas the other defects.

In other words, each political official controls complete political hegemony when she unilaterally defects (i.e., she proposes an unreasonable constitutional change catered to her particular comprehensive doctrine while rejecting any constitutional amendment proposed by the other political official), whereas the other cooperates (i.e., she approves any constitutional change proposed by the other political official while proposing a reasonable constitutional amendment herself); achieves mutual justice when both cooperate (i.e., when both political officials propose and approve constitutional amendments in accordance with the political conception of justice); suffers political instability when both defect (i.e., when both political officials reject the other party's proposal and try to impose their particular comprehensive doctrine throughout society); and suffers political subjugation when she unilaterally cooperates, whereas the other defects (i.e., when she proposes constitutional amendments in accordance with the political conception of justice that get rejected, and unilaterally approves the proposal made by the other party who intends to impose her particular comprehensive doctrine). Figure 1 represents the social outcomes that may be generated in our well-ordered society by the combination of each political official's actions.

Each public official's preferences are represented by a payoff function $u_i : O \times T \to \mathbb{R}$, which is a function of

the political outcome and the public official's type—that is, public official *i*'s payoff is denoted $u_i(o, t)$. Specifically:

$$u_i\,(\text{Hegemony}, u) = H$$
$$u_i\,(\text{Justice}, u) = 1$$
$$u_i\,(\text{Instability}, u) = 0$$
$$u_i\,(\text{Subjugation}, u) = -S$$
$$u_i\,(\text{Justice}, r) = J$$
$$u_i\,(\text{Hegemony}, r) = 1$$
$$u_i\,(\text{Instability}, r) = 0$$
$$u_i\,(\text{Subjugation}, r) = -S$$

H, J, and S are positive real numbers such that $S > J$, $H > 1 > 0$. The payoffs reflect the different preferences of the reasonable (type *r*) and the unreasonable (type *u*) public officials. We can see that although the reasonable types most prefer to achieve mutual justice, the unreasonable types most prefer to win complete political hegemony, which would allow them to create a perfectionist state in accordance with their moral/religious beliefs. That is, although the reasonable types most prefer to live under a well-ordered liberal democratic society—the characteristic feature of which is what Rawls called "reasonable pluralism"—the unreasonable types wish to monopolize state power to coerce everybody to believe their particular comprehensive moral/religious doctrine.

The payoffs also show that, regardless of one's type, the cost of being politically subjugated is greater than what one can achieve positively either from political hegemony or mutual justice (i.e., $|-S| > J$, $H > 0$). This is in accordance with the fundamental importance that Rawls had put on securing each individual's freedom, as well as what Rawls called "highest-order interests." According to Rawls, not only do people have preferences over specific policies that are formed on the basis of their particular conceptions of the good, but they also have what may be called a *meta-preference* for their conceptions of the good to be formed and affirmed under conditions that are noncoercive and free (1971/1999, 131–32). When people are politically subjugated by another group of people, this type of freedom—the freedom to choose and revise one's particular ends, which Rawls believes to be of a fundamental importance—gets destroyed. This is the reasoning behind assuming $|-S| > J$, $H > 0$.

Here is a brief description of how the game is played. Let $\pi \in (0, 1]$ denote the proportion of unreasonable types who have entered our well-ordered society. At the start of the game, public official 1 is informed about his type, which public official 2 does not know. Instead, based on the proportion of unreasonable types who reside in our well-ordered society, public official 2 believes that the probability that public official 1 is unreasonable is π, and the probability that public official 1 is reasonable is $1 - \pi$. Both public officials then simultaneously decide to play either *c* (i.e., cooperate) or *d* (i.e., defect). The payoffs for each public official are generated, and the game ends.

FIGURE 1 Social Outcomes in a Well-Ordered Society

1 \ 2	c	d
c	Justice, Justice	Subjugation, Hegemony
d	Hegemony, Subjugation	Instability, Instability

Let us call this game "the game of the well-ordered society under crisis"; it is the well-ordered society under "crisis," as it presumes the existence of noncooperating, unreasonable types. The normal (strategic) forms of the game are presented, first described in outcomes (Figure 2) and next described in payoffs (Figure 3).

The dotted lines indicate the respective information sets for each public official. The game shows that each type of public official 1 knows his own type as well as the fact that public official 2 is reasonable, whereas public official 2, while knowing that she is reasonable, believes that she is interacting with an unreasonable political official (i.e., $1u$) with probability $\pi > 0$ and a reasonable public official (i.e., $1r$) with probability $1 - \pi$. Note that the reasonable types are modeled as *conditional cooperators*—they prefer to cooperate when their counterpart cooperates, and they prefer to defect when their counterpart defects. By contrast, the unreasonable types are modeled as full stop *noncompliers*; for them, defecting (i.e., playing *d*) is a strictly dominant strategy.

Note that our baseline model is a *one-shot game*. This is to model that our well-ordered society is facing a crisis—going through a constitutional change that would, once made, be practically irrevocable for an unforeseeable future. We will follow K&S and assume that the game is played between two high-ranked public officials (2016, 719), whose mutual agreement is sufficient to change the constitution in a particular direction. Hence, the game does not represent a situation in which ordinary citizens are trying to decide whether or not to comply with the specific requirements of their society's law (e.g., pay taxes); rather, the situation concerns two high-ranked public officials deliberating about how to reorganize the basic structure of their society by amending the

FIGURE 2 The Game of the Well-Ordered Society under Crisis (Outcomes)

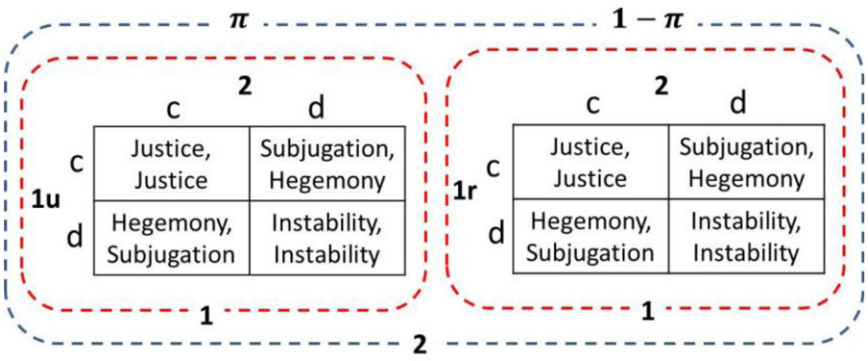
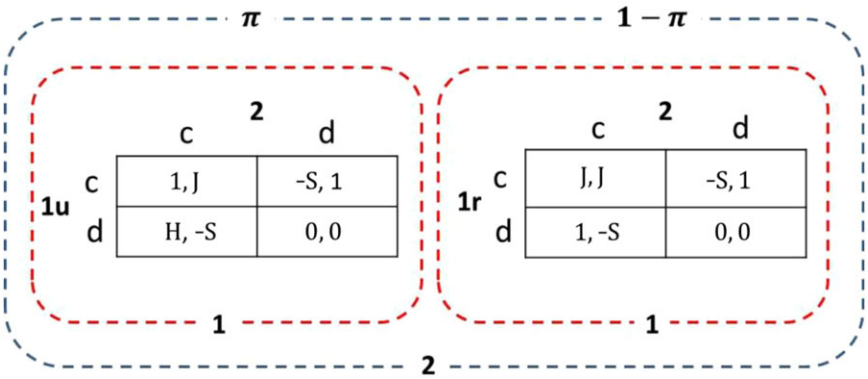


FIGURE 3 The Game of the Well-Ordered Society under Crisis (Payoffs)



constitution. Hence, the game represents a situation in which penal sanctions (Rawls 1997/1999, 505) cannot effectively provide assurance that the other political official will not try to change the constitution in an unreasonable way.

Now, given the gravity of the situation, how would the strategic interactions of the two public officials unfold? As the proportion of unreasonable types in our well-ordered society increases, the probability that one's cooperation would be taken advantage of, resulting in one's political subjugation, increases. Given that the reasonable types wish to avoid such political subjugations, we would expect there to be a threshold of the proportion of unreasonable types over which the reasonable types would judge that it would be optimal for them to simply defect. We are curious to know *the minimum* threshold of unreasonable types over which it would *never* be sequentially rational for the reasonable types to cooperate. The following proposition identifies this minimum threshold and further claims that it will always be below majority.

Proposition 1. *Let* $\underline{\pi}$ *denote the minimum threshold of unreasonable types such that whenever* $\pi > \underline{\pi}$*, it would never be sequentially rational for the reasonable types (i.e., type r) to cooperate (i.e., play c) with any positive probability. Then* $\underline{\pi} = \frac{J-1}{J+S-1} < \frac{1}{2}$*—that is, such a threshold will always be below majority.*[8]

According to Proposition 1, the minimum threshold of unreasonable types over which it is never sequentially rational for the reasonable types to cooperate with any positive probability is $\underline{\pi} = \frac{J-1}{J+S-1}$. How big this value $\underline{\pi} = \frac{J-1}{J+S-1}$ will be will depend on the two parameters S and J—that is, the disvalue of political subjugation as well the value of justice. The specific values of these two parameters will depend on the two components of one's sense of justice that Rawls assumes: (a) the willingness to achieve social justice by reciprocating cooperation to other cooperators and (b) the unwillingness to be politically subjugated. Component (a) determines the value of J, whereas

[8]See Appendix for proof.

component (b) determines the value of S. Hence, as one's sense of justice becomes stronger, both parameters, S and J, will likely increase. Whenever one's distaste for political subjugation (i.e., component b) becomes stronger relative to one's subjective value for mutual justice (i.e., component a), the threshold $\underline{\pi} = \frac{J-1}{J+S-1}$ decreases, resulting in a higher likelihood of social instability. Conversely, whenever one's subjective value for mutual justice (i.e., component a) becomes stronger relative to one's distaste for political subjugation (i.e., component b), the threshold $\underline{\pi} = \frac{J-1}{J+S-1}$ increases (as $\frac{\partial \underline{\pi}}{\partial J} = \frac{S}{(J+S-1)^2} > 0$). However, as proved in Proposition 1, the threshold $\underline{\pi} = \frac{J-1}{J+S-1}$ is guaranteed to be below $\frac{1}{2}$. In other words, in the model, the proportion of unreasonable people that is sufficient to completely destabilize the well-ordered society is *strictly less than majority*; or to say the same thing differently, the well-ordered society can completely destabilize even when we have a strict majority of reasonable public officials.[9] With Proposition 1, we can now easily prove a sufficient condition for complete destabilization of our well-ordered society.

Corollary of Proposition 1. *Suppose* $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$. *Then the strategy profile* $((s_{1u}, s_{1r}), s_2) = ((d, d), d)$ *is the unique Bayesian Nash equilibrium (BNE) of the game of well-ordered society under crisis.*[10]

Corollary of Proposition 1 claims that, given that the proportion of unreasonable types that have entered into the well-ordered society is greater than some endogenously determined threshold, which is strictly less than one-half, the well-ordered society will necessarily destabilize. The main mechanism that drives this result is *uncertainty*. If the reasonable types were able to identify each other and avoid interacting with unreasonable types, mutual justice would have been an equilibrium of the game. The reason why this is not possible is because the reasonable types are unable to distinguish other reasonable types from the unreasonable types.

First Extension: The Well-Ordered Society under Crisis with Public Reason

As we have seen from the previous section, the main reason why the well-ordered society destabilizes under crisis is due to the inability of the reasonable types to properly distinguish themselves from the unreasonable types. Then the question is this: Would the reasonable political officials be able to identify one another and coordinate theirs ways toward mutual justice if they were able to signal their types to other reasonable people *through the use of public reason* in such a way that reliably indicates their firm commitment to the political conception of justice? This section explores this possibility.

For this purpose, we extend the model and allow public official 1 to communicate (i.e., signal) his type to public official 2 via using public reason. Note that Rawls's main stance on what sort of reasons citizens in a liberal democratic society are allowed to use when engaging in public deliberation has changed throughout his career. Initially, Rawls had adopted what many call the "exclusive view," according to which reasons that stem from one's comprehensive moral or religious doctrines should not enter into public deliberations, full stop. However, later in "The Idea of Public Reason Revisited," Rawls has relaxed such a requirement and endorsed what he calls "the wide view," according to which

> reasonable comprehensive doctrines, religious or nonreligious, may be introduced in public political discussion at any time, provided that in due course proper political reasons—and not reasons given solely by comprehensive doctrines—are presented that are sufficient to support whatever the comprehensive doctrines introduced are said to support. (Rawls 1993/2005, 462)

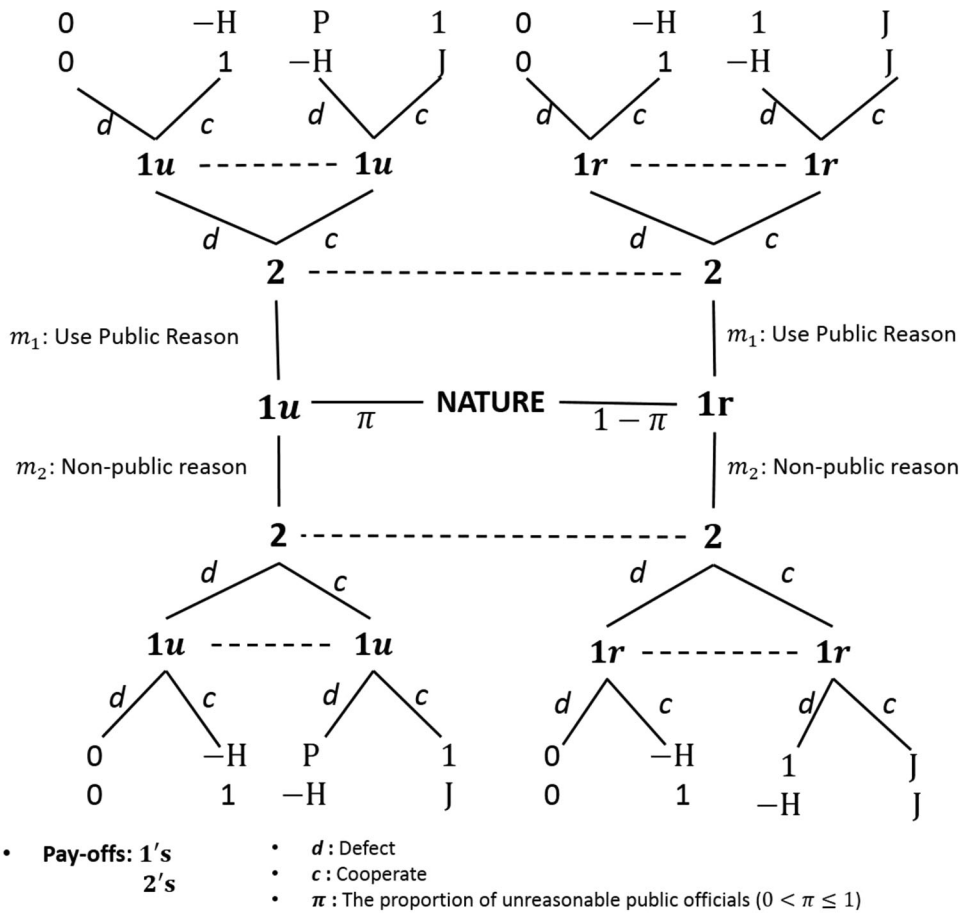
Rawls called the qualification made in the "provided that" clause "the proviso." Weithman interprets "the proviso" as allowing "ordinary citizens to rely on their comprehensive doctrines without adducing public reasons in support of their positions, so long as their doing so does not lead others to doubt that they acknowledge the authority of the public conception of justice" (2015, 88). The situation that we will consider in this section is a situation in which the sort of doubt that Weithman mentions has arisen. More specifically, in our situation, suppose public official 2, who is reasonable, is starting to suspect whether public official 1 is showing proper allegiance toward the political conception of justice undergirding their liberal democratic society—that is, public official 2 is starting to doubt whether public official 1 is reasonable or just rational but unreasonable.

Here is how the game is played. At first, NATURE determines a probability distribution over the two types of public official 1, where π is the probability that public official 1 is unreasonable and $1 - \pi$ is the probability that public official 1 is reasonable. After public official

[9] Chung (2019) shows that the threshold of unreasonable type over which the well-ordered society completely destabilizes can be made arbitrarily low whenever such an intrusion destroys the common knowledge precondition of the well-ordered society.

[10] See Appendix for proof.

FIGURE 4 The Well-Ordered Society under Crisis with Public Reason



1 ascertains whether he is unreasonable or reasonable, he must send a public message to public official 2. Let $M = \{m_1, m_2\}$ be the set of messages that public official 1 can send to public official 2, where m_1 denotes a public message that relies solely on public reasons and m_2 denotes a public message that contains nonpublic reasons. After public official 1 sends a public message, the two public officials simultaneously decide whether to cooperate or defect. Payoffs are generated as before. The extensive form of the game is depicted in Figure 4.

Now, in the model, we can see that neither message (i.e., m_1 or m_2) has any effect on any of the public officials' payoffs; payoffs are entirely determined by one's type as well as the combination of actions performed by each public official. Hence, the messages in the model meet the formal definition of *cheap talk* in game theory.

One question we might ask at this point is whether modeling the use of public reason as cheap talk is plausible. I believe so. This is due to Rawls's own commitment to the full publicity condition of the political conception of justice. According to Rawls,

> The full justification of the public conception of justice ... [is] to be publicly known, or better, at least to be *publicly available*. This weaker condition (that full justification be available) allows for the possibility that some will not want to carry philosophical reflection about political life so far, and certainly no one is required to. But if citizens wish to, the full justification is present in the public culture, reflected in its system of law and political institutions, and in the main historical traditions of their interpretation. (Rawls 1993/2005, 67; emphasis added)

Note how low Rawls is setting the bar for the general use of public reason. One does not have to philosophically reflect about it in order to use it in public life. In a well-ordered society, public reasons are fully embedded within the political culture here and there and are readily publicly available. This means that one does not have to exert that much effort to use public reason in public deliberation. Rawls had deliberately made public reason cheap and

affordable in the well-ordered society. Hence, modeling public reason as cheap talk is perfectly in alignment with Rawls's general intentions. The two major results of the model are as follows.

Proposition 2. *There exists no Perfect Bayesian Equilibrium (PBE) in which the reasonable types can separate themselves from the unreasonable types and successfully achieve justice with any positive probability.*[11]

Proposition 3. *Suppose* $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$. *Then the only equilibrium outcome is instability.*[12]

The formal analyses of the model as well as the formal proofs of Propositions 2 and 3 are relegated to the Appendix. Here, let us just briefly discuss informally what these two propositions imply for our main discussion.

What Proposition 2 is essentially saying is that even if the reasonable public officials try to coordinate with other reasonable public officials by relying on public reason, this, in general, will not work, as the unreasonable public officials have every incentive to also send public messages in the veneer of public reason whenever it can induce cooperation from other reasonable public officials, which they can fully exploit to control complete political hegemony. Remember this is possible since using public reason in a well-ordered society (thanks to Rawls's own publicity condition) is virtually "for free." Proposition 2 shows that the use of public reason will not be able to distinguish the unreasonable public officials from the reasonable public officials.[13]

Proposition 3 goes a little further. It says that whenever the proportion of unreasonable public officials in government exceeds the threshold that was determined in the previous section, not only will we not be able to distinguish the reasonable political officials from the unreasonable ones by using public reason, but it will also be impossible to restore social justice with *any* positive probability no matter how low. In others, despite the use of public reason, the well-ordered society completely destabilizes.

Both Propositions 2 and 3 together imply that, as a device for mutual coordination, public reason is useless. It confirms the criticisms raised by Gaus (2011), Thrasher and Vallier (2015), and K&S (2016)—namely, that public reason, by being merely "cheap talk," will not be able to solve the problem of political stability in the well-ordered society.[14]

Second Extension: The Well-Ordered Society under Crisis with Convergence Discourse

We have seen that the main reason why the reasonable public officials were unable to use public reason as a coordinating device was because it was too cheap; unreasonable public officials who are not actually committed to the political conception of justice can readily afford to use public reason to induce other public officials' cooperation, which they can exploit. K&S's solution was to make public deliberation "costly" via convergence discourse. The thought was that doing so would make it hard for the unreasonable public officials to mimic the signals of the reasonable public officials.

Unlike the consensus framework of public justification, the convergence framework of public justification allows citizens to appeal to reasons stemming from their comprehensive doctrines. As already explained, on the face of it, it is unclear how this would make public deliberation costly: If everybody is allowed to appeal to her own comprehensive doctrine, would this not actually make engaging in public deliberation easier and, hence, cheaper? According to K&S, this is not so because if a public official, who is, say, a Christian, wishes to convince another public official, who is, say, a Hindu, to endorse some policy *p* under convergence discourse, unlike the consensus framework, the first public official would

[11] See Appendix for proof.

[12] See Appendix for proof.

[13] Here is a real-world example: In 2014, Lee Seok-ki, a member of South Korea's left-wing United Progressive Party (통합진보당), who was also a lawmaker of the South Korean National Assembly at that time, was arrested and sentenced to 12 years in prison for having links to North Korea and allegedly plotting an armed rebellion against the South Korean government (see https://edition.cnn.com/2014/02/17/world/asia/south-korea-lee-rebellion-plot-conviction/index.html). Lee was a strong follower of Juche philosophy (주체사상/主體思想), the founding philosophy of North Korea and its official state ideology, which combines Marxist–Leninist communism with a strong form of nationalism with the deification of "the Great Leader" (referring to Kim Il-sung and his "noble" bloodline). Lee, like many other Juche philosophy followers, intended to aid North Korea in reunifying Korea and ultimately establish a perfectionist state based on the teachings of Juche philosophy. Lee and his party supporters have long claimed to abolish the National Security Act (on the basis of which he was arrested) primarily on *liberal grounds*—namely, that the act violates core liberal democratic values such as freedom of thought/expression/association and is therefore unconstitutional. We now know that Lee's liberal rhetoric was simply a mask to disguise his real intentions to help create a perfectionist state under North Korean rule on the Korean Peninsula.

[14] This is in accordance with other game-theoretic work in political science. For instance, Austen-Smith (1990) shows that "debate" (modeled as cheap talk) cannot elicit any new information that would change the final outcome of a political process.

have to expend a considerable amount of effort to learn the details of Hinduism to provide Hindu-based reasons in support of policy p to fully convince the second public official. "That [the first public official] would be willing to incur such a cost indicates that she is serious about achieving the [mutual justice] outcome" (K&S 2016, 724–5).

Let us now extend our baseline model to incorporate this logic of convergence discourse. Consider again our well-ordered society under crisis. Suppose $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$; that is, the proportion of unreasonable public officials in our well-ordered society exceeds the threshold that is sufficient to destabilize liberal democratic political order. Let there be $N \in \mathbb{N}$ comprehensive doctrines in our well-ordered society and let public official 2, who is reasonable, be committed to one of the N comprehensive doctrines.

Here is the timing of the model. At first, NATURE gives a probability distribution $(\pi,\ 1-\pi)$ to the two types of public official 1—that is, the unreasonable type (i.e., type u) and the reasonable type (i.e., type r). Each type of public official 1, knowing his type, decides on the amount of effort to exert to learn the various comprehensive doctrines in his society. Let $e_u \geq 0$ be the amount of effort expended by the unreasonable type (i.e., type u) of public official 1 and $e_r \geq 0$ be the amount of effort expended by the reasonable type (i.e., type r) of public official 1. Given effort level $e \geq 0$, the probability that each type of public official 1 learns public official 2's comprehensive doctrine is $\min\{1, \frac{e}{N}\}$. Also, exerting $e \geq 0$ amount of effort costs the public official the same amount $e \geq 0$ that gets subtracted from his final payoff. Note that each type of public official 1 can learn public official 2's comprehensive doctrine for certain by expending $e \geq N$ amount of effort. As N is the number of comprehensive doctrines in society, we can see that when there are more comprehensive doctrines in society, the amount of effort that would be required for each type of public official 1 to learn public official 2's comprehensive doctrine for certain increases. To put it another way, learning public official 2's comprehensive doctrine becomes costlier as society becomes more diverse. This is in line with what will be referred to as K&S's "the more diversity the better (MDB) thesis," according to which

> The signal from Row's convergence discourse is significantly more costly in the large diversity case than in the small diversity case . . . . [T]he 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*—the too cheap talk problem. (K&S 2016, 726; emphasis in original)

We will soon provide a more in-depth formal analysis of K&S's MDB thesis after we characterize the equilibrium of the model.

To continue, we assume that public official 2 does not directly observe the specific effort levels expended by each type of public official 1. Instead, once each type of public official 1 exerts effort, public official 2 receives a signal: a "good" signal (g) if a given type of public official 1 has successfully learned public official 2's comprehensive doctrine, and a "bad" signal (b) otherwise. We assume that the signal that public official 2 receives is also private information; that is, which specific signal public official 2 has received is unobservable to each type of public official 1. After public official 2 receives a signal, both types of public official 1 and public official 2 simultaneously decide whether to cooperate or to defect. Once they do, payoffs are generated as in our baseline model.

Again, our solution concept is perfect Bayesian equilibrium (PBE). We have seen from Proposition 1 that, given $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$, the well-ordered society under crisis will never succeed in restoring social justice with any positive probability. We have seen in the previous section that this result is unchanged even if we allow the public officials to use public reason as a public communicative signal. We now wish to examine whether convergent discourse can solve our current predicament as K&S (2016) claim.

It is worth noting that the current extension incorporating the logic of convergence discourse has multiple PBEs, many of which do no better than public reason. Hence, we are interested in finding a PBE in which convergence discourse is successful. Specifically, we are interested in finding a PBE in which public official 2 cooperates if public official 1 successfully learns public official 2's comprehensive doctrine (i.e., if public official 2 receives the "good" (g) signal) and defects otherwise (i.e., if public official 2 receives the "bad" (b) signal); the reasonable type of public official 1 exerts a positive amount of effort to learn public official 2's comprehensive doctrine; and the well-ordered society under crisis achieves mutual justice with positive probability. Let us call such a PBE a *convergence discourse equilibrium* (CDE):

> **Convergence Discourse Equilibrium (CDE)**: A CDE in a well-ordered society under crisis is a PBE in which public official 2 cooperates after observing signal g and defects after observing signal b; the reasonable type (i.e., type r) of public official 1 exerts positive effort (i.e., $e_r > 0$);

and the outcome "mutual justice" is achieved with positive probability.

One natural thing to ask is whether such a CDE will always exist in the well-ordered society under crisis once we introduce convergence discourse. The answer is a disappointing "no."

Proposition 4. *Suppose* $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$. *Then there exists no convergence discourse equilibrium if the value of political hegemony* (H) *is sufficiently large, specifically, when* $H > N$.[15]

What Proposition 4 is saying is this. When the unreasonable type of public official 1's desire to achieve political hegemony is strong relative to the number of comprehensive doctrines in our well-ordered society under crisis (i.e., $\frac{H}{N} > 1$), the optimal level of effort is $e_u = N$; that is, the unreasonable type of public official 1 will learn every comprehensive doctrine in his society. When this happens, public official 2's strategy to cooperate after observing that public official 1 has learned public official 2's comprehensive doctrine is no longer sustainable. This is so because the fact that public official 2 received a "good" (g) signal no longer gives any reason for public official 2 to believe that her counterpart is more likely to be reasonable than before, as the good signal could have equally been produced by an unreasonable counterpart. It is obvious that there can be no CDE in this case, and, as a result, our well-ordered society destabilizes completely just as it was the case when the public officials relied on public reason.

This shows that introducing convergence discourse does *not* somehow magically solve the problem that public reason failed to solve. Based on our analysis, a necessary condition for a CDE to exist is $\frac{H}{N} \leq 1 \Rightarrow H \leq N$. That is, either the number of comprehensive doctrines must be sufficiently large or the unreasonable type's desire to achieve political hegemony must be sufficiently weak; whenever this fails, there is no CDE.

So, for the purpose of our argument, let us assume that $H \leq N$. That is, let us assume that relative to the number of comprehensive doctrines that coexist in our well-ordered society under crisis, conditions are relatively favorable in the sense that the unreasonable types' desire to achieve political hegemony is not too strong. Can convergence discourse now perform the role that K&S hopes it will perform by restoring social justice and political stability by serving as a costly signal? To answer this question in the affirmative, we must show that a CDE exists. Fortunately, a CDE does exist, which the next proposition demonstrates:

Proposition 5. *Suppose* $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$ *and suppose* $H \leq N$. *Then the following assessment constitutes a convergence discourse equilibrium:*

- Public official 1's (type u) strategy
  $s_{1u} = \begin{cases} e_u = 0 \\ \text{D(defect)} \end{cases}$;
- Public official 1's (type r) strategy
  $s_{1r} = \begin{cases} e_r = \frac{NS}{J+S-1} \\ \text{C (cooperate)} \end{cases}$;
- Public official 2's strategy
  $s_2 = \begin{cases} \text{C (cooperate) after observing } g \\ \text{D (defect) after observing b} \end{cases}$;
- $1u$'s beliefs: believes that he has generated signal b with probability 1;
- $1r$'s beliefs: believes that he has generated signal g with probability $\frac{S}{J+S-1}$ and signal b with probability $1 - \frac{S}{J+S-1}$;
- Public official 2's beliefs
  $= \begin{cases} \text{after observing } g, \text{ believes that } 1 = 1u \\ \text{for certain after observing } b, \\ \text{believes } 1 = 1u \text{ with probability} \\ \frac{\pi}{\pi+(1-\pi)(1-\frac{S}{J+S-1})} \end{cases}$.[16]

The CDE that we have just found shows that it may be possible, *under relatively favorable conditions*, for convergence discourse to solve the problem that public reason failed to solve—namely, the problem of restoring social justice and maintaining political stability when the well-ordered society faces a crisis. In the previous section, we saw that whenever the proportion of unreasonable types who have intruded into our well-ordered society exceeds $\underline{\pi} = \frac{J-1}{J+S-1}$, relying on public reason was a sure way to completely destabilize our well-ordered society. Proposition 5 shows that, given that the political ambitions of the unreasonable types who are threatening the political stability of the well-ordered society are not too high, relying on convergence discourse *may* help restore and maintain social justice and political order by serving as a costly signal that the unreasonable types cannot be reasonably expected to mimic. This partially confirms that K&S's (2016) 'solution' can actually be a solution to the problem that a well-ordered society under crisis creates. However, there are reasons to think that we should not be too optimistic.

The biggest problem is that even when conditions are relatively favorable in the sense that the political ambitions of the unreasonable types are not too high (i.e., even when $H \leq N$), the CDE that we found in Proposition 5 is *not* the *only equilibrium* of the model; as a matter

[15] See Appendix for proof.

[16] See Appendix for proof.

of fact, there is *a continuum* of (that is, infinitely many) equilibria in which our well-ordered society under crisis will *destabilize completely* despite the use of convergence discourse. The next proposition characterizes a *family* of such PBEs in which instability is the unique equilibrium outcome of our well-ordered society under crisis despite the use of convergence discourse:

Proposition 6. *Suppose* $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$ *and suppose* $H \leq N$. *Then the following characterizes a family of PBEs (that are not CDEs) in which our well-ordered society completely destabilizes:*

- Public official 1's (type u) strategy $s_{1u} = \begin{cases} e_u = 0 \\ \text{D (defect)} \end{cases}$;
- Public official 1's (type r) strategy $s_{1r} = \begin{cases} e_r = 0 \\ \text{D(defect)} \end{cases}$;
- Public official 2's strategy $s_2 = \begin{cases} \text{D(defect) after observing } g \\ \text{D (defect) after observing } b \end{cases}$;
- $1u$'s beliefs: believes that he or she has generated signal b with probability 1;
- $1r$'s beliefs: believes that he or she has generated signal $\underline{b}$ with probability 1;
- Public official 2's beliefs $= \begin{cases} \text{after observing } g, \text{ believes that } 1 = 1u \\ \text{with probability } \lambda (\text{where } \lambda > \underline{\pi}) \\ \text{after observing } b, \\ \text{believes } 1 = 1u \text{ with probability } \pi \end{cases}$.[17]

What Proposition 6 shows is that even under relatively favorable conditions in which the political ambitions of the unreasonable types are not too strong, our well-ordered society under crisis can very well destabilize completely despite the use of convergence discourse.

What is more is that even when convergence discourse does work successfully and our well-ordered society under crisis is in the good CDE, we still cannot say that it is a complete solution to restore social justice and political stability in the well-ordered society. In the CDE discovered in Proposition 5, the probability that our well-ordered society under crisis will be able to restore political justice by using convergence discourse is $P^* = (1-\pi)(\frac{S}{J+S-1})$. (This is the probability that the reasonable type of public official 1 succeeds in learning public official 2's comprehensive doctrine.) Given our parameters (viz. $\frac{J}{J+S} < \pi \leq 1,\ S > J > 1$), P^* is guaranteed to be below 1. That is, even when using convergence discourse as a costly signal successfully, there will always be some positive probability that our well-ordered society under crisis will fail to restore social justice and destabilize nonetheless. So even when our well-ordered society under crisis is in the good convergence discourse equilibrium, using convergence discourse as a costly signal does not completely solve the problem; it only solves the problem *sometimes*.

So how often is "sometimes"? To examine this, let us do some comparative statics. We can see that $P^* = (1-\pi)(\frac{S}{J+S-1})$ (i.e., the probability that our well-ordered society under crisis restores and maintains social justice and political order by using convergence discourse as a costly signal) is determined by three parameters: π (the proportion of unreasonable types who have intruded into society); J (how much the reasonable types value mutual justice); and S (the magnitude of the disvalue of being politically subjugated.) Consequently, we may write $P^*(\pi, J, S)$. To see how $P^*(\pi, J, S)$ responds to a change in each parameter, let us take the partial derivatives of $P^*(\pi, J, S)$ with respect to each parameter:

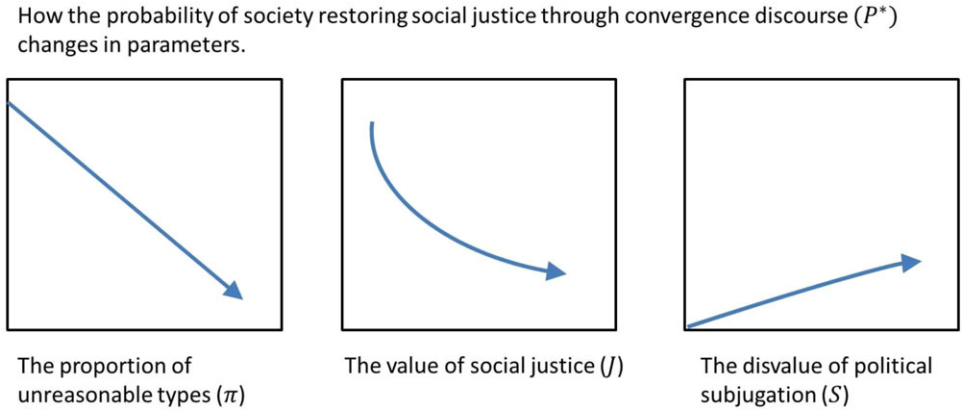
- $\frac{\partial P^*(\pi,J,S)}{\partial \pi} = \frac{-S}{J+S-1} < 0$
- $\frac{\partial P^*(\pi,J,S)}{\partial J} = \frac{-(1-\pi)S}{(J+S-1)^2} < 0$
- $\frac{\partial P^*(\pi,J,S)}{\partial S} = \frac{-(1-\pi)(1-J)}{(J+S-1)^2} > 0$

The qualitative interpretations of these partial derivatives would amount to this. The probability that the well-ordered society that faces a crisis successfully restores social justice and maintains political stability by using convergence discourse as a costly signal (a) *decreases* as more unreasonable types intrude into society (i.e., when π increases), (b) *decreases* as the reasonable types value mutual justice more strongly (i.e., when J increases), and (c) *increases* as the reasonable types disvalue political subjugation more strongly (i.e., when S increases.) Figure 5 summarizes this graphically.

Component (a) and (c) seem natural; one would expect that it would be harder to restore social justice even with costly signaling if there are more unreasonable types in society, and one would also expect that the reasonable types will expend more effort to learn the comprehensive doctrines in their society to signal (to other reasonable types) their genuine intention to cooperate the more they fear being politically subjugated by the unreasonable types, which results in a higher probability of the convergence discourse solution succeeding. What one might find surprising is component (b)—namely, that the probability that our well-ordered society under crisis successfully restores social justice and maintains political stability by using convergence discourse decreases as reasonable people start to value mutual justice more strongly. This means that, unlike Rawls's prediction, the educational effects of

[17] See Appendix for proof.

FIGURE 5 Comparative Statics: Graph of P^*



the well-ordered society (Rawls 2001, 185) might actually engender the seeds for its very own destabilization whenever the well-ordered society faces a crisis.

Let us now examine what I have called K&S's MDB (the more diversity the better) thesis. According to the MDB thesis, the more diverse the society (i.e., as the number of comprehensive doctrines increases), the easier it will be for convergence discourse to solve the problem of political stability in a well-ordered society facing a crisis. The MDB thesis is partially supported by the model.

As we already know, a necessary condition for our CDE to exist is $H \leq N$. That is, the unreasonable type's desire to achieve political hegemony must be sufficiently weak *relative to* the number of comprehensive doctrines that coexist in the well-ordered society. We can see that this condition will be easier to be met as N gets larger. Whenever this condition is met, expending zero effort (i.e., $e_u = 0$) will be the only effort level that would be sequentially rational for the unreasonable types, given public official 2's strategy to cooperate after observing a "good" (g) signal and to defect after observing a "bad" (b) signal. This means that whenever N is sufficiently large (i.e., whenever the well-ordered society is sufficiently diverse), the unreasonable types will simply give up learning any of their society's comprehensive doctrines. In contrast, in the CDE, the reasonable types will always exert positive effort to learn their society's comprehensive doctrines. So whenever a reasonable official finds that her counterpart has sufficiently learned her comprehensive doctrine, she will know for sure that her counterpart is also reasonable, and, from this, both political officials can cooperate toward mutual justice. To this extent, the MDB thesis is true.

However, there is a flip side to this. As we have seen, whenever $H > N$, that is, whenever the unreasonable types' desire to achieve political hegemony becomes sufficiently large, the unreasonable types will learn every single comprehensive doctrine in the well-ordered society they infiltrated. Hence, a reasonable type will no longer be able to distinguish the reasonable types from an unreasonable type simply by observing whether or not her counterpart has successfully learned her comprehensive doctrine. As demonstrated from Proposition 4, whenever this happens, the well-ordered society will necessarily destabilize, and convergence discourse will no longer serve as a viable solution to the problem.

Hence, we may say that promoting diversity is a double-edged sword. It may make it easier for a well-ordered society under crisis to restore social justice and political stability via the means of convergence discourse, but it may also destroy this very possibility if it goes too far to promote religious fundamentalists and extremist comprehensive doctrines (whose desires to achieve political hegemony are unusually strong). To the extent that diversity promotes extremist views in society, the MDB thesis is false. This is a reason why we might want to contain diversity in a reasonable way. This sort of thing might be what Rawls had in mind when he claimed that "political liberalism takes for granted not simply pluralism but the fact of *reasonable pluralism*" (Rawls 1993/2005, xviii; emphasis added).

Concluding Remarks

Any theory of political stability must take into consideration the potential disruptive effects of noncompliers. This is because if a political system is prone to destabilize with the introduction of a small number of unreasonable public officials who have the power to restructure the political system, then this shows that the political stability initially achieved was too fragile to serve as a basis for

modern liberal political institutions. In this sense, Rawls's theory of political stability (i.e., "stability for the right reasons") is incomplete, as it simply assumes the existence of noncompliers away. Thrasher and Vallier (2018) have recently argued that this type of *closed-ness* is one of the key defects of Rawls's conception of the well-ordered society and have proposed an alternate model of a liberal democratic society, which they call "the open society," according to which constitutional rules, while remaining stable, still "preserve the social conditions that foster experimentation, while leaving room in legal and institutional rules for innovation and change" (398).

In this article, we have examined whether the two prominent solutions offered in the public reason liberalism literature—namely, using public reason or using convergence discourse—can restore social justice and maintain political stability when the well-ordered society faces a crisis, particularly when the proportion of unreasonable public officials who have intruded into the well-ordered society exceeds some endogenously determined threshold. The formal analyses offered in this article shows that using public reason fails completely, whereas using convergent discourse, although slightly better, has its own critical limitations. Despite Rawls's high hopes, it seems that achieving "stability for the right reasons" and successfully maintaining it is not something that can be achieved so easily.

There are many ways in which the main models presented in this article can be further extended. One possible way is to divide the unreasonable types into two subcategories (e.g., unreasonable type A and unreasonable type B) and add a third public official who is assumed to be a given unreasonable type. By assuming that the unreasonable types of the same type can credibly identify and cooperate with each other, we may be able to examine the extent to which coalitional assortment can affect the proper operations of public reason and convergence discourse. Another possible way to extend the model is to introduce additional periods and assume that the status quo of the well-ordered society under crisis worsens the longer it takes for the well-ordered society to restore justice. In such a setup, the reasonable types will face a trade-off; they can cooperate in the current period while facing a risk of turning the well-ordered society into a perfectionist state or they can stall the political process for another round with the hope of uncovering the other political official's type while the crisis worsens. These and other possible extensions are left for future research.

Although the primary focus of this article was to examine, through formal game-theoretic analysis, the effectiveness of the two prominent solutions that have been offered to restore political order in Rawls's well-ordered society, the results of the paper have broader implications for the theory of deliberative democracy more generally. The crucial idea of the theory of deliberative democracy is that a given policy proposal is legitimate to the extent that it can be justified by reasons that those who are trying to find fair terms of social cooperation among free and equal persons cannot reasonably reject (Cohen 1997, 73; Gutman and Thomson 2004, 3). However, what the formal results of this article show is that whether or not a given policy proposal is well supported by good reasons may not in itself be sufficient for its stable public endorsement and implementation if there exists reasonable suspicion among the decision-making parties about the genuine intentions of their decision-making counterparts.

I would like to end by noting that the application of formal game-theoretic analysis has proved to be fruitful in identifying the specific conditions and institutional requirements under which democratic deliberation can successfully achieve the many normative aims that democratic theorists have hoped. It has been shown that the success of democratic deliberation in achieving better outcomes through information sharing can heavily depend on the postdeliberation voting rule (Austen-Smith and Feddersen 2006; Coughlan 2000; Mathis 2011); and contrary to the received wisdom among deliberative democratic theorists, informational efficiency may even be better served by unequal and asymmetric standing of the speakers with respect to the deliberative procedure (Hafer and Landa 2007). More generally, game-theoretic analyses have made it clear that different deliberative environments can provide different incentive structures that may either promote or hinder successful deliberation among strategic actors (Landa and Meirowitz 2009).

Any normative political theory presumes an understanding of the positive processes of its proposed institutional arrangement. As such, game-theoretic analysis, by offering a better understanding of the positive processes of different kinds of deliberative institutions, can greatly help the normative theorizing of democratic theorists by showing what can and what cannot be realistically expected from the practical implementation of their normative political theories. This is why there should be more fruitful collaborations among formal/positive and normative political theorists.

References

Austen-Smith, David. 1990. "Information Transmission in Debate." *American Journal of Political Science* 34(1): 124–52.

Austen-Smith, David, and Timothy J. Feddersen. 2006. "Deliberation, Preference Uncertainty, and Voting Rules." *American Political Science Review* 100(2): 209–17.

Chung, Hun. 2019. "The Instability of John Rawls's 'Stability for the Right Reasons.'" *Episteme* 16(1): 1–17.

Chung, Hun. 2019. "The Impossibility of Liberal Rights in a Diverse World." *Economics and Philosophy* 35(1): 1–27.

Chung, Hun, and Brian Kogelmann. 2018. "Diversity and Rights: a Social Choice-Theoretic Analysis of the Possibility of Public Reason." *Synthese*. https://doi.org/10.1007/s11229-018-1737-4

Cohen, Joshua. 1997. "Deliberation and Democratic Legitimacy." In *Deliberative Democracy: Essays on Reason and Politics*, ed. James Bohman and William Rehg. Cambridge, MA: MIT Press: 67–92.

Coughlan, Peter. 2000. "In Defense of Unanimous Jury Verdicts: Mistrials, Communication and Strategic Voting." *American Political Science Review* 94(2): 375–93.

Gaus, Gerald. 2011. "A Tale of Two Sets: Public Reason in Equilibrium." *Public Affairs Quarterly* 25(4): 305–25.

Gutman, Amy, and Dennis Thomson. 2004. *Why Deliberative Democracy?* Princeton, NJ: Princeton University Press.

Hadfield, Gillian, and Stephen Macedo. 2012. "Rational Reasonableness: Toward a Positive Theory of Public Reason." *Law & Ethics of Human Rights* 6(1): 1–46.

Hadfield, Gillian, and Barry Weingast. 2012. "What Is Law? A Coordination Model of the Characteristics of Legal Order." *Journal of Legal Analysis* 4(2): 471–514.

Hafer, Catherine, and Dimitri Landa. 2007. "Deliberation as Self-Discovery and Institutions for Political Speech." *Journal of Theoretical Politics* 19(3): 329–60.

Kogelmann, Brian, and Stephen Stich. 2016."When Public Reason Fails Us: Convergence Discourse as Blood Oath." *American Political Science Review* 110(4): 717–30.

Landa, Dimitri, and Adam Meirowitz. 2009. "Game Theory, Information, and Deliberative Democracy." *American Journal of Political Science* 53(2): 427–44.

Mathis, Jérôme. 2011. "Deliberation with Evidence." *American Political Science Review* 105(3): 516–29.

Muldoon, Ryan. 2016. *Social Contract Theory for a Diverse World: Beyond Tolerance*. New York: Routledge.

Rawls, John. 1997/1999. *A Theory of Justice*. Revised edition. Cambridge, MA: Harvard University Press.

Rawls, John. 1993/2005. *Political Liberalism*. Expanded edition. New York: Columbia University Press.

Rawls, John. 2001. *Justice as Fairness: A Restatement*. Cambridge, MA: Belknap Press of Harvard University Press.

Thrasher, John, and Kevin Vallier. 2015. "The Fragility of Consensus: Public Reason, Diversity and Stability." *European Journal of Philosophy* 23(4): 933–54.

Thrasher, John, and Kevin Vallier. 2018. "Political Stability in the Open Society." *American Journal of Political Science* 62(2): 398–409.

Weithman, Paul. 2015. "Inclusivism, Stability, and Assurance." In *Rawls and Religion*, ed. Tome Bailey and Valentina Gentile. New York: Columbia University Press, 75–98.

Appendix: Proofs and Technical Details

Proposition 1. *Let $\underline{\pi}$ denote the minimum threshold of unreasonable types such that whenever $\pi > \underline{\pi}$, it would never be sequentially rational for the reasonable types (i.e. type r) to cooperate (i.e. play c) with any positive probability. Then, $\underline{\pi} = \frac{J-1}{J+S-1} < \frac{1}{2}$ – that is, such a threshold will always be below majority.*

Proof. For public official $1u$, defecting (i.e. playing d) strictly dominates cooperating (i.e. playing c). Hence, $s_{1u} = d$ is the only sequentially rational strategy for public official $1u$. Let $\sigma_{1r}(c) = p$ – that is, let p denote the probability that public official $1r$ cooperates (i.e. play c). Then, for public official 2, the expected payoff of defecting (i.e. playing d) is:

$$\begin{aligned} EU_2(d) &= \pi \cdot 0 + (1-\pi)(p \cdot 1 + (1-p) \cdot 0) \\ &= (1-\pi)\, p \end{aligned}$$

Similarly, the expected payoff of cooperating (i.e. playing c) is:

$$\begin{aligned} EU_2(c) &= \pi \cdot (-S) + (1-\pi)(p \cdot J + (1-p) \cdot (-S)) \\ &= -\pi (J+S)\, p + (J+S)\, p - S \end{aligned}$$

It is sequentially rational for public official 2 to defect if:

$$\begin{gathered} EU_2(d) > EU_2(c) \\ (1-\pi)\, p > -\pi (J+S)\, p + (J+S)\, p - S \\ (J+S-1)\, p\pi > (J+S-1)\, p - S \\ \pi > \frac{(J+S-1)\, p - S}{(J+S-1)\, p} \cdots (*) \end{gathered}$$

Let $\pi(p) = \frac{(J+S-1)p-S}{(J+S-1)p}$. Then, we have $\pi'(p) = \frac{S}{(J+S-1)p^2} > 0$. Hence, the right-hand side of the inequality $(*)$, i.e. $\pi(p)$ is a strictly increasing function of $p \in [0, 1]$, which takes its maximum value $\pi(1) = \frac{J-1}{(J+S-1)}$. Set $\underline{\pi} = \frac{J-1}{J+S-1}$. Then, whenever $\pi > \underline{\pi}$, it is never sequentially rational for public official 2 to cooperate regardless of the value of $p \in [0, 1]$. Hence, $s_2 = d$ whenever $\pi > \underline{\pi}$. Knowing that public official 2 will optimally defect whenever $\pi > \underline{\pi}$, it is sequentially rational for the public official $1r$ to defect as well. Hence, $s_{1r} = d$ whenever $\pi > \underline{\pi}$. Hence, whenever $\pi > \underline{\pi}$, it is never sequentially rational for any reasonable types to cooperate with positive probability. To show that $\underline{\pi} = \frac{J-1}{J+S-1}$ is *the minimum* threshold of unreasonable types over which it is never sequentially rational for the reasonable types to cooperate, it suffices to show that, for any lower threshold, it would be possible for there to exist a positive probability with which playing c would be sequentially rational

for public official 2. So, pick any $\epsilon > 0$, and suppose we set $\underline{\pi} = \frac{J-1}{J+S-1} - \epsilon$. Suppose $\pi > \underline{\pi} = \frac{J-1}{J+S-1} - \epsilon$. From $(*)$, it would be sequentially rational for public official 2 to cooperate if and only if $\pi \leq \frac{(J+S-1)p-S}{(J+S-1)}$. Suppose $p = 1$. Then, whenever $\underline{\pi} = \frac{J-1}{J+S-1} - \epsilon < \pi \leq \frac{J-1}{J+S-1}$, it would be sequentially rational for public official 2 to cooperate. Hence, $\forall \epsilon > 0$, $\underline{\pi} = \frac{J-1}{J+S-1} - \epsilon$ cannot be the minimum threshold. I conclude that $\underline{\pi} = \frac{J-1}{J+S-1}$ is the minimum threshold as required. Finally, I claim that $\underline{\pi} = \frac{J-1}{J+S-1} < \frac{1}{2}$. Otherwise, $\frac{J-1}{J+S-1} \geq \frac{1}{2} \Rightarrow 2(J-1) \geq J+S-1 \Rightarrow J - S \geq 1$, which is impossible as $S > J$ by assumption. ■

Corollary of Proposition 1. *Suppose* $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$. *Then, the strategy profile,* $((s_{1u}, s_{1r}), s_2) = ((d, d), d)$ *is the unique Bayesian Nash equilibrium (BNE) of the game of well-ordered society under crisis.*

Proof. For the unreasonable type of public official 1, playing d strictly dominates playing c. Therefore, $s_{1u} = d$ is the only sequentially rational strategy unreasonable type of public official 1. By the proof of the Proposition 1, given $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$, $s_2 = s_{1r} = d$ are the only set of sequentially rational strategies for public official 2 and the reasonable type of public official 1. Therefore, the strategy profile $((s_{1u}, s_{1r}), s_2) = ((d, d), d)$ is the unique Bayesian Nash equilibrium (BNE) of the well-ordered society under crisis. ■

Technical Details of Section 3: The Game of Well-ordered Society under Crisis with Public Reason

To analyze the game in Figure 4 more formally, we would need to introduce some more notations. We can see that the game contains a total of 15 nodes. Let $X = \{x_0, x_1, \ldots, x_{14}\}$ be the set of all nodes in the game. Each node in X can be identified with a sequence of actions played up to that point. Specifically, let:

$$x_0 = (\phi)$$
$$x_1 = (u)$$
$$x_2 = (r)$$
$$x_3 = (u, m_1)$$
$$x_4 = (r, m_1)$$
$$x_5 = (u, m_2)$$
$$x_6 = (r, m_2)$$
$$x_7 = (u, m_1, d)$$
$$x_8 = (u, m_1, c)$$
$$x_9 = (r, m_1, d)$$
$$x_{10} = (r, m_1, c)$$
$$x_{11} = (u, m_2, d)$$
$$x_{12} = (u, m_2, c)$$
$$x_{13} = (r, m_2, d)$$
$$x_{14} = (r, m_2, c)$$

For instance, $x_0 = (\phi)$ denotes the initial node at which NATURE determines the probability distribution over the two types of public official 1, and $x_7 = (g, m_1, d)$ denotes the node that is reached after NATURE determines public official 1 as a reasonable type, public official 1 sends the message using public reason (i.e. sends m_1), and, then, public official 2 attacks.

The set $I = X \backslash \{x_0\}$ can be partitioned into eight information sets such that $I = \{I_1, \ldots, I_8\}$ where $I_1 = \{x_1\}$, $I_2 = \{x_2\}$, $I_3 = \{x_3, x_4\}$, $I_4 = \{x_5, x_6\}$, $I_5 = \{x_7, x_8\}$, $I_6 = \{x_9, x_{10}\}$, $I_7 = \{x_{11}, x_{12}\}$, $I_8 = \{x_{13}, x_{14}\}$.

Let $\theta(I_i)$ denote the public official whose turn it is to play at information set I_i $(\mathrm{i} = 1, \ldots, 8)$. So, we have $\theta(I_1) = \theta(I_5) = \theta(I_7) = 1u$, $\theta(I_2) = \theta(I_6) = \theta(I_8) = 1r$, and $\theta(I_3) = \theta(I_4) = 2$. Let $A(I_i)$ denote the set of actions available to $\theta(I_i)$ at I_i. So, $A(I_1) = A(I_2) = M = \{m_1, m_2\}$ and $A(I_i) = A = \{d, c\}$ for $i = 3, 4, \ldots, 8$.

A *behavioral strategy* of public official $\theta(I_i)$ at information set I_i is $\sigma_{\theta(I_i)} : I_i \rightarrow \Delta A(I_i)$ – that is, a behavioral strategy of public official $\theta(I_i)$ at information set I_i is a probability distribution on the actions available at the information set I_i. For $\mathrm{a}_1 \in A(I_i)$, let $\sigma_{\theta(I_i)}(a_1 | I_i)$ denote the probability with which public official $\theta(I_i)$ plays a_1 at information set I_i. Let $\sigma = (\sigma_{1m}, \sigma_{1r}, \sigma_2)$ be the profile of behavioral strategies for public officials 1 and 2.

A *system of beliefs* $\mu = (\mu_{I_1}, \ldots, \mu_{I_8})$ is a profile of probability distributions, one for each information set I_i, such that for all $I_i \in I$, $\mu_{I_i}(x) \in [0, 1]$ for all $x \in I_i$ and $\sum_{x \in I_i} \mu_{I_i}(x) = 1$.

Let $u_{\theta(I_i)}(\sigma | I_i, \mu_{I_i})$ denote the payoff of public official $\theta(I_i)$ at information set $\mathrm{I_I}$ with beliefs μ_{I_i} and strategies σ.

An *assessment* (σ, μ) is a pair consisting of a profile of behavioral strategies and a system of beliefs. Our solution concept is *Perfect Bayesian Equilibrium* (PBE).

An assessment (σ, μ) is a PBE if:

1. It is *sequentially rational*–that is, for all $I_i \in I$ and all $\sigma'_{\theta(I_i)}$,

$$u_{\theta(I_i)}(\sigma | I_i, \mu_{I_i}) \geq u_{\theta(I_i)}(\sigma'_{\theta(I_i)}, \sigma_{-\theta(I_i)} | I_i, \mu_{I_i})$$

And

1. *All beliefs obey Bayes' rule whenever possible*– that is, for all $I_i \in I$ such that $\Pr[I_i|\sigma] > 0$,

$$\mu_{Ii}(x) = \frac{\Pr[x|\sigma]}{\Pr[I_i|\sigma]} \text{ for all } x \in I_i.$$

We are now ready to state and prove Propositions 2 and 3:

Proposition 2. *There exists no PBE in which the reasonable types can separate themselves from the unreasonable types and successfully achieve justice with positive probability.*

Proof. Suppose not. Then, there exists a separating PBE in which 1r and 2 play *c* with positive probability. Let the assessment (σ, μ) be any such separating PBE. Since (σ, μ) is separating, we must have: $\sigma_{1u}(I_1) \neq \sigma_{1r}(I_2)$. Without loss of generality, suppose $\sigma_{1u}(m_1|I_1) = 0$ and $\sigma_{1u}(m_1|I_2) = 1$ (i.e. suppose that public official 1 sends the message using public reason when he/she is a reasonable type, and sends the message using non-public reason when he/she is an unreasonable type.) Since, in a PBE, beliefs are assigned according to Bayes' rule whenever possible, we must have $\mu_{I_3}(x_3) = 0$ and $\mu_{I_4}(x_5) = 1$ (i.e. given that public official 2 receives a message encoded in public reason, public official 2 knows for sure that public official 1 is a reasonable type) Since defecting strictly dominates cooperating for the unreasonable types, we must, by sequential rationality, have $\sigma_{1u}(d|I_5) = \sigma_{1u}(d|I_7) = 1$. Given m_2 (i.e. given that public official 2 receives the message in non-public reason), defecting gives public official 2 a payoff of 0, while cooperating gives public official 2 a payoff of $-S$. Therefore, by sequential rationality, $\sigma_2(d|I_4) = 1$. This results in a payoff of 0 for $1u$ (i.e. the unreasonable type of public official 1). Let $\sigma_2(c|I_3) = q$. By assumption, $q \in (0, 1)$ (i.e. once receiving the message in public reason, public official 2 will cooperate with positive probability.) If $1u$ in I_1 deviated to $\sigma_{1u}(m_1|I_1) = 1$ and played defect afterwards, then this will give him/her a payoff of: $q \cdot H + (1-q) \cdot 0 = q \cdot H > 0$. So, this is a profitable deviation. So, $\sigma_{1u}(m_1|I_1) = 0$ is not sequentially rational. This contradicts that $\sigma_{1u}(m_1|I_1) = 0$ is part of a PBE. ■

Proposition 3. *Suppose* $> \underline{\pi} = \frac{J-1}{J+S-1}$. *Then, the only equilibrium outcome is instability.*

Proof. Suppose not. Then, there exists a PBE in which the equilibrium outcome is not instability. This can happen only if both 1r and 2 play c with positive probability in some information set reached with positive probability. Let the assessment (σ, μ) be any such PBE.

First, by Proposition 2, (σ, μ) cannot be a *separating equilibrium.*

Second, I claim that (σ, μ) cannot be a *pooling equilibrium.* For, suppose not. Then, either $\sigma_{1u}(m_1|I_1) = \sigma_{1r}(m_1|I_2) = 1$ or $\sigma_{1u}(m_1|I_1) = \sigma_{1r}(m_1|I_2) = 0$. Without loss of generality, suppose $\sigma_{1u}(m_1|I_1) = \sigma_{1r}(m_1|I_2) = 1$. Then, $\mu_{I_3}(x_3) = \pi$ and $\mu_{I_3}(x_4) = 1 - \pi$. We can see that the game has essentially been reduced to the baseline model that we have seen in section 2. Since $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$, by Proposition 1, it is never sequentially rational for any reasonable type to cooperate with any positive probability, which contradicts our assumption that both 1r and 2 public official 2 cooperates with positive probability.

Therefore, the only possibility left is for (σ, μ) to be a *semi-separating equilibrium.* Let $\sigma_{1u}(m_1|I_1) = w \in (0, 1)$ and $\sigma_{1r}(m_1|I_2) = z \in (0, 1)$. Since (σ, μ) is semi-separating, the information sets I_3 and I_4 are reached with positive probability. By assumption, public official 2 needs to cooperate with positive probability in at least one of these information sets. Without loss of generality, suppose that public official 2 cooperates with positive probability in I_3; that is, $\sigma_2(c|I_3) \in (0, 1]$. In order for playing c with positive probability to be sequentially rational for public official 2 in I_3, we must have $\mu_{I_3}(x_3) \leq \frac{J}{S+J}$, or else, playing d will strictly dominate playing c, and, hence, playing c with positive probability will not be sequentially rational. Since S > J, $\frac{J}{S+J} < \frac{1}{2}$. Hence, in order for $\mu_{I_3}(x_3) \leq \frac{J}{S+J} < \frac{1}{2}$, we must have w < z. This implies $\sigma_{1u}(m_2|I_1) = (1-w) > (1-z) = \sigma_{1r}(m_2|I_2)$. By Bayes' rule, this implies that $\mu_{I_4}(x_5) > \frac{1}{2} > \frac{J}{S+J}$. Given such beliefs, playing d strictly dominates playing c for public official 2 in I_4. Therefore, sequential rationality requires $\sigma_2(d|I_4) = 1$. Given $\sigma_2(d|I_4) = 1$, sequential rationality requires $\sigma_{1u}(d|I_7) = \sigma_{1r}(d|I_8) = 1$. (In other words, the bottom portion of the game results in Instability.) Such a combination of strategies gives $1u$ a payoff of:

$$w[\sigma_2(c|I_3)(H) + (1 - \sigma_2(c|I_3)) \cdot 0] + (1-w)(0)$$
$$= w \cdot \sigma_2(c|I_3)(H).$$

If $1u$ deviated to $\sigma_{1u}(m_1|I_1) = 1$, then his/her payoff would become: $\sigma_2(c|I_3)(H)$. Since $w \in (0, 1)$, this is a strictly profitable deviation. Therefore, $\sigma_{1u}(m_1|I_1) = w$ is not sequentially rational. This contradicts that $\sigma_{1u}(m_1|I_1) = w$ is part of a PBE. This shows that there cannot be a semi-separating PBE in which 1r and 2 cooperate with positive probability in the equilibrium path. We have exhausted all cases. Therefore, we conclude that there exists no PBE in which the equilibrium outcome is not Instability. ■

Proposition 4. *Suppose* $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$. *Then, there exists no CDE if the value of political hegemony (H) is sufficiently large (i.e.* $H > N$.*)*

Proof. Assume $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$ and $H > N$. Suppose, contrary to the claim, that there exists a CDE. By the definition of a CDE, public official 2 cooperates after observing signal g and defects after observing signal b. Call this strategy s_2 (i.e. $s_2(g) = c$ and $s_2(b) = d$). Now, a strategy for each type of public official 1's consists of *two parts*: the first part specifies the amount of effort to learn public official 2's comprehensive doctrine, and the second part specifies whether to cooperate or defect. We know that defection strictly dominates cooperation for the unreasonable type of public official 1(i.e. $1u$) in the second stage of the game. Hence, the only strategy that could be supported by sequential rationality for $1u$ after public official 2 observes a signal (either g or b) is to defect. (So, $s_{1u} = \text{D}$ will form the second part of the unreasonable type of public official 1's strategy.) Now, let us deduce the optimal amount of effort $e_u \geq 0$ for $1u$, which will constitute the first part of $1u$'s strategy. Note that any effort level greater than N is strictly dominated by effort level $e_u = N$ (as expending a greater effort level than N would not increase the probability of learning public official 2's comprehensive doctrine and would only incur a greater cost.) So, in any equilibrium, we must have $e_u \in [0, N]$. Given public official 2's strategy s_2 (i.e. $s_2(g) = c$ and $s_2(b) = d$), the expected payoff of expending $e_u \geq 0$ amount of effort for the unreasonable type of public official 1 is:

$$EU_{1u}(e_u|s_2) = \frac{e_u}{N}(H - e_r) + \left(1 - \frac{e_u}{N}\right)(-e_r)$$
$$= e_u\left(\frac{H}{N} - 1\right) \cdots (*)$$

Since $H > N$ by assumption, we have $\frac{H}{N} - 1 > 0$, which means that the expected payoff for $1u$ (i.e. equation $(*)$) is increasing in e_u, i.e. the amount of effort expended. Hence, the optimal level of effort for $1u$ becomes $e_u = N$. By expending $e_u = N$ amount of effort, $1u$ generates a 'good signal (g)' with probability 1. But then, after observing the g, player 2 believes (after Bayesian updating) that the probability that s/he is facing an unreasonable type is at least $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$. By Proposition 1, whenever $\pi > \underline{\pi}$, it is never sequentially rational for the reasonable types (i.e. type r) to cooperate (i.e. play c) with any positive probability, which contradicts our initial assumption that $s_2(g) = c$ is part of a CDE and hence sequentially rational. ■

Proposition 5. *Suppose $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$ and suppose $H \leq N$. Then the following assessment constitutes a CDE:*

- *Public official 1 (type u)'s strategy*
  $s_{1u} = \begin{cases} e_u = 0 \\ \text{D(defect)} \end{cases}$;
- *Public official 1 (type r)'s strategy*
  $s_{1r} = \begin{cases} e_r = \frac{NS}{J+S-1} \\ \text{C(cooperate)} \end{cases}$;
- *Public official 2's strategy*
  $s_2 = \begin{cases} \text{C(cooperate) after observing } g \\ \text{D(defect) after observing } b \end{cases}$;
- *$1u$'s beliefs: believe that s/he has generated signal 'b' with probability 1;*
- *$1r$'s beliefs: believe that s/he has generated signal 'g' with probability $\frac{S}{J+S-1}$ and signal 'b' with probability $1 - \frac{S}{J+S-1}$.*
- *Public official 2's beliefs*
  $= \begin{cases} \text{after observing } g, \text{ believes that } 1 = 1u \\ \text{for sure after observing } b, \text{ believes} \\ 1 = 1u \text{ with probability} \\ \frac{\pi}{\pi + (1-\pi)(1-\frac{S}{J+S-1})} \end{cases}$

Proof. First, let us verify that the stated assessment is a PBE. To do so, let us verify that $1u$'s strategy s_{1u} is sequentially rational. For $1u$, defection strictly dominates cooperation in the second stage of the game. Hence, the only strategy that could be supported by sequential rationality for $1u$ after public official 2 observes any signal is to defect. Hence, in any equilibrium, we must have $s_{1u} = D$. Given public official 2's stated strategy s_2 (i.e. $s_2(g) = c$ and $s_2(b) = d$), the expected payoff of expending $e_u \geq 0$ amount of effort for $1u$ is:

$$EU_{1u}(e_u|s_2) = \frac{e_u}{N}(H - e_r) + \left(1 - \frac{e_u}{N}\right)(-e_r)$$
$$= e_u\left(\frac{H}{N} - 1\right) \cdots (*)$$

Since $H \leq N$, we have $\frac{H}{N} - 1 < 0$, and, hence, the optimal level of effort for $1u$ is $e_u = 0$. This verifies that $1u$'s stated strategy s_{1u} is sequentially rational.

Next, let us verify that the reasonable type of public official $1(1r)$'s stated strategy s_{1r} is sequentially rational. Note that any effort level greater than N is strictly dominated by effort level $e_r = N$, as expending a greater effort level than N would not increase the probability of learning public official 2's comprehensive doctrine and would only incur a greater cost. So, suppose the reasonable type of public official 1 exerts effort level $e_r \in [0, N]$. Given public official 2's stated strategy s_2, after expending e_r, the expected payoff of cooperating for $1r$ is:

$$EU_{1r}(c|e_r, s_2)$$
$$= \frac{e_r}{N}(J - e_r) + \left(1 - \frac{e_r}{N}\right)(-S - e_r) \cdots (1)$$

The expected payoff of defecting is:

$$EU_{1r}(d|e_r, s_2) = \frac{e_r}{N}(1-e_r) + \left(1-\frac{e_r}{N}\right)(0-e_r)\cdots(2)$$

Cooperating is sequentially rational if and only if $(1) \geq (2)$, which, after solving for e_r, implies $e_r \geq \frac{NS}{J+S-1}$. So, for the reasonable type of public official 1, the optimal level of effort to expend to learn public official 2's comprehensive doctrine while making cooperation sequentially rational is $e_r = \frac{NS}{J+S-1}$. We need this level of effort to be less or equal to N. Note $e_r = \frac{NS}{J+S-1} \leq N \Rightarrow \frac{S}{J+S-1} \leq 1$, which is always satisfied as $J > 1$. This verifies that $1r$'s stated strategy s_{1r} is sequentially rational.

Let us now verify that public official 2's stated strategy s_2 is sequentially rational. When $1r$ exerts $e_r = \frac{NS}{J+S-1}$, the probability that s/he will learn public official 2's comprehensive doctrine, and, thereby, produce a 'good' (g) signal, is: $\frac{e_r}{N} = \frac{S}{J+S-1}$. Since $1u$ will exert zero effort (i.e. $e_u = 0$), public official 2 will observe a 'good' (g) signal exclusively from the reasonable type of public official 1's efforts. Hence, the total probability that public official 2 will observe a 'good' (g) signal in our well-ordered society under crisis is: $(1-\pi)(\frac{S}{J+S-1})$. Knowing that a 'good' (g) signal can only be produced by $1r$'s efforts, after observing a 'good' (g) signal, public official 2 will believe, after Bayesian updating, that s/he is interacting with $1r$ for sure. Given that $1r$ cooperates, it is sequentially rational for public official 2 to cooperate as well. Hence, public official 2's optimal strategy after observing the 'good' signal is to cooperate: $s_2^*(g) = c$.

Now, suppose that public official 2 has received a 'bad' (b) signal. There are two cases in which this can happen: (i) when public official 2 is interacting with $1u$ (which occurs with probability: π), or (ii) when the $1r$ fails to learn public official 2's comprehensive doctrine despite his/her efforts to learn (which occurs with probability: $(1-\pi)(1-\frac{S}{J+S-1})$). So, by applying Bayes' rule, the conditional probability that public official 2 will be interacting with an unreasonable type of public official 1 *given* that public official 2 observes a 'bad' (b) signal is: $\frac{\pi}{\pi+(1-\pi)(1-\frac{S}{J+S-1})}$, which is greater than π, which, in our model, is assumed to be greater than $\frac{J-1}{J+S-1}$. By Proposition 1, we know that whenever public official 2 believes that the proportion of unreasonable public officials exceeds $\frac{J-1}{J+S-1}$, the only sequentially rational strategy is for public official 2 to defect; hence, $s_2^*(b) = d$.

Hence, we have verified that our assumed strategy for public official 2 is sequentially rational given both types of public official 1's strategies and public official 2's beliefs. All information sets are reached with positive probability, and the beliefs assigned in the assessment is in accordance with Bayesian updating. Hence, the assessment is a PBE.

As a final step, we verify that the proposed assessment is also a CDE. note that $s_2(g) = c$, $s_2(b) = d$; $e_r = \frac{NS}{J+S-1} > 0$; and mutual justice is achieved with probability $P^* = (1-\pi)(\frac{S}{J+S-1}) > 0$. ■

Proposition 6. *Suppose* $\pi > \underline{\pi} = \frac{J-1}{J+S-1}$ *and suppose* $H \leq N$. *Then, the following characterizes a family of PBEs in which our well-ordered society completely destabilizes:*

- *Public official 1 (type u)'s strategy*
  $s_{1u} = \begin{cases} e_u = 0 \\ \text{D(defect)} \end{cases}$;
- *Public official 1 (type r)'s strategy*
  $s_{1r} = \begin{cases} e_r = 0 \\ \text{D(defect)} \end{cases}$;
- *Public official 2's strategy*
  $s_2 = \begin{cases} \text{D(defect)after observing } g \\ \text{D(defect) after observing } b \end{cases}$;
- $1u$*'s beliefs: believe that s/he has generated signal 'b' with probability 1;*
- $1r$*'s beliefs: believe that s/he has generated signal 'b' with probability 1;*
- *Public official 2's beliefs*
  $= \begin{cases} \text{after observing } g, \text{ believes that } 1 = 1u \\ \text{with probability } \lambda (\text{where } \lambda > \underline{\pi}) \\ \text{after observing } b, \text{ believes } 1 = 1u \text{ with} \\ \text{probability } \pi \end{cases}$

Proof. Given $s_2(g) = s_2(b) = d$, the stated strategies of each type of public official 1 generate the following payoffs: $U_{1u}(s_{1u}|s_2) = U_{1r}(s_{1r}|s_2) = 0$. For each type $i = u, r$ of public official 1, deviating to either $e_i > 0$ or C (cooperate) (or both) will result in a payoff of: $-e_i < 0$ or $-S < 0$ or $-S - e_i < 0$. Hence, neither type of public official 1 has any incentive to deviate to a different strategy. Now, let us check that public official 2's strategy s_2 is sequentially rational. After public official 2 receives a signal (either good or bad), the game is reduced to the baseline model of a well-ordered society under crisis depicted in figure 3 of section 2. By Proposition 1, it is never sequentially rational for public official 2 to cooperate with any positive probability whenever s/he believes that the probability that s/he is facing an unreasonable type is greater than $\underline{\pi} = \frac{J-1}{J+S-1}$. Hence, given public official 2's assigned beliefs, $s_2(g) = s_2(b) = d$ is sequentially rational. We now check that the assigned beliefs obey Bayes' rule in all information sets reached in the equilibrium path. Since neither type of public official 1 exerts any effort, the probability that a 'good signal(g)' is produced is

zero. Hence, each type of public official 1 should believe that s/he has produced a 'bad signal(b)' with probability one. Next, after observing a b, public official 2 should believe through Bayesian updating that s/he is interacting with $1u$ with probability: $\frac{\pi(1)}{\pi(1)+(1-\pi)(1)} = \pi$. Observing g is off-the-equilibrium-path. Hence, we are allowed to assign any beliefs for public official 2 in this information set. Let λ be public official 2's belief that s/he is facing $1u$ after observing a g. Then, by Proposition 1, any $\lambda > \underline{\pi} = \frac{J-1}{J+S-1}$ will render $s_2(g) = d$ sequentially rational. We conclude that the stated assessment characterizes a family of PBEs that are not CDEs. ■