Episteme, 16, 1 (2019) 1–17 © Cambridge University Press
doi:10.1017/epi.2017.14

THE INSTABILITY OF JOHN RAWLS'S "STABILITY FOR THE RIGHT REASONS"

HUN CHUNG
hunchung1980@gmail.com

ABSTRACT

John Rawls's most mature notion of political order is "stability for the right reasons." Stability for the right reasons is the kind of political order that Rawls hoped a well-ordered society could ideally achieve. In this paper, I demonstrate through the tools of modern game theory, the *instability* of "stability for the right reasons." Specifically, I will show that a well-ordered society can completely destabilize by the introduction of an arbitrarily small number of non-compliers whenever individuals fail to achieve full common knowledge ever so slightly.

1. COMMON KNOWLEDGE AS A PRECONDITION FOR STABILITY FOR THE RIGHT REASONS

John Rawls wrote *A Theory of Justice* (*TJ*) (Rawls 1971/1999) in order to present a superior alternative to utilitarianism. With the publication of *TJ*, Rawls thought that his battle against utilitarianism has concluded with his victory. By the time Rawls wrote *Political Liberalism* (*PL*), his aims had changed (Rawls 1993/2005). He now wanted to show how it is possible that "there may exist over time a stable and just society of free and equal citizens profoundly divided by reasonable religious, philosophical, and moral doctrines" (*PL*: xxv). For this purpose, Rawls repackaged his theory of justice, known as "justice as fairness", as a free-standing *political* conception of justice, which is justified, first and foremost, on the basis of public reasons that can appeal to everybody as free and equal citizens regardless of his/her particular comprehensive religious, philosophical beliefs (*PL*: 450). A second layer of stability was achieved by what Rawls called an "overlapping consensus" under which each citizen finds additional supporting reasons for the political conception of justice from his/her particular comprehensive religious, philosophical doctrine (*PL*: 134). Rawls called the type of political stability so achieved as "stability for the right reasons" (*PL*: 458–60) and contrasted it with the type of stability achieved by a mere balance of powers, a stability which Rawls called *modus vivendi*.

In short, stability for the right reasons requires three things. First, it requires everybody to fully endorse the same political conception of justice solely on the basis of shared public reasons. Second, it requires everybody to be fully motivated, from his/her sense of justice (and not merely from a fear of punishment or self-interest), to abide by the requirements of public justice conditional upon others' likely motivation. And, third, it requires these facts to be known as a matter of *common knowledge*.

The third condition – namely that everybody endorses as well as everybody knows that everybody endorses the same political conception of justice as a matter of common

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

knowledge – is a crucial component that undergirds Rawls's notion of stability for the right reasons. Stability for the right reasons is a notion of political stability which Rawls thought to be achieved and successfully maintained in a "well-ordered society." In defining a well-ordered society, Rawls explains that he is considering

> a society in which everyone accepts, and knows that everyone else accepts, the very same political conception of justice (and so the same principles of political justice.). Moreover, this knowledge is *mutually recognized*: that is, people know everything they would know if their acceptance of those principles were a matter of public agreement. (Rawls 2001: 8, emphasis added)

That is, a defining characteristic of a well-ordered society is that the fact that everybody endorses the same political conception of justice is "mutually recognized." Of course, it is not entirely clear whether, here, Rawls was requiring full common knowledge (in its full technical sense) or merely mutual knowledge as a defining characteristic of a well-ordered society. The main difference between the two is that mutual knowledge is a significantly weaker epistemic requirement than full common knowledge – to explain, for any two people to have mutual knowledge, it suffices for each to know that the other is a full endorser of the political conception of justice regulating the basic structure of his/her society. That is, mutual knowledge simply requires each person to know the other person's *type*. Common knowledge, in addition, requires such mutual knowledge to fully extend to the other person's (high-order) *beliefs* – that is, not only does one have to know that the other is a full endorser of the political conception of justice, one has to know that the other knows that one knows that the other is a full endorser of the political conception of justice, and so on.

Although Rawls was not entirely clear on this matter, it is not too difficult to understand (from the perspective of rational choice theory) that mere mutual knowledge (i.e. knowing that other people are reasonable cooperators) would have been too weak of a precondition to sustain stability for the right reasons in a well-ordered society. This is because, in Rawls's well-ordered society, reasonable people (i.e. those who have successfully developed their requisite senses of justice) are *conditional* cooperators – viz. they are the type of people who wish to cooperate and honor the requirements of the political conception of justice *only when* others are willing to cooperate and contribute their fair share as well. Merely knowing that my counterpart is such a reasonable type is insufficient to *assure* me that my counterpart will actually do his/her part of the agreement – this is because my counterpart might decide to renege if she suspects (either justifiably or unjustifiably) that I will not cooperate, and this might be because either she believes that I am an unreasonable type or she believes that *I* believe that she is an unreasonable type, and so on. Brian Skyrms (2004) nicely illustrates the key features of such a situation, when he writes

> There is an incentive to avoid unilateral deviation, but, for example, if you expect me to deviate, you might believe you would be better off deviating as well. And if I believe that you have such beliefs, I may expect you to deviate and by virtue of such expectations deviate myself. (Skyrms 2004: 51)

The problem, here, does not stem from the two people being stuck in a Prisoner's Dilemma type of situation in which both persons have a dominant strategy to defect

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

based on considerations of self-interest; both people, being reasonable, and having successfully internalized the requisite senses of justice, consider mutual cooperation as their most preferred outcome. The problem, rather, stems from *the failure of mutual assurance*. We can see that Rawls was clearly aware of this problem when he writes

> since even with a sense of justice men's compliance with a cooperative venture is predicated on the belief that others will do their part, citizens may be tempted to avoid making a contribution *when they believe, or with reason suspect*, that others are not making theirs. These tendencies arising from *apprehensions* about the faithfulness of others lead to instability of the second kind. (Rawls 1971/1999: 296, emphasis added)

What Rawls seems to be saying here is that merely knowing each other's *type* (i.e. mutual knowledge) is insufficient to achieve stability for the right reasons in a well-ordered society; what is further needed is to know each other's *beliefs* (i.e. common knowledge). This is exactly the kind of consideration that led David Lewis (2002) to require common knowledge as a precondition for the successful establishment and maintenance of a *convention* (Lewis 2002: 52–60). If Rawls wishes to regard stability for the right reasons as a successfully established and maintained convention in his well-ordered society, Rawls would have to follow Lewis. Hence, we can say that, even though Rawls himself was not entirely explicit on this matter, the fact that everybody endorses the same public conception of justice as a matter of (not simply mutual knowledge but also) common knowledge is something that Rawls should have proposed more explicitly as a defining characteristic of his well-ordered society. Furthermore, since a well-ordered society is a precondition for the very achievement and maintenance of stability for the right reasons, it follows that common knowledge itself is a precondition for the very achievement and maintenance of stability for the right reasons.

One thing to note: Rawls explains that, in a well-ordered society, "(nearly) everyone strictly complies with, and so abides by, the principles of justice" (Rawls 2001: 13) conditional upon others' likely compliance. I would like to point to Rawls's use of the adjective "nearly" to qualify "everyone" in the previous quoted sentence. By doing so, Rawls is actually allowing the possibility of the existence of a small number of non-compliers (i.e. those who are rational but unreasonable) in a well-ordered society.

There are others who have also noted that Rawls's well-ordered society allows the existence of a small number of rational but unreasonable non-compliers. For instance, in a recent manuscript titled "Consensus on What? Convergence for What?: Four Models of Political Liberalism", Gerald Gaus points out that the "overlapping consensus on justice as fairness does not ... require that all reasonable comprehensive doctrines participate in this equilibrium " (Gaus 2016: 9). Instead, what Rawls's overlapping consensus effectively requires is for a "plurality of opposing comprehensive doctrines" and a "substantial majority of politically active citizens" to participate (Gaus 2016: 9). This is because, according to Gaus, for Rawls, "stability is a population-level trait, and so only the overall population of reasonable doctrines must be in a social normative equilibrium" (Gaus 2016: 9). In other words, according to Gaus's interpretation, the overlapping consensus, which is an integral part of stability for the right reasons, requires, not everybody, but only a substantial majority of reasonable people to participate for its overall success. This leaves room for there to be a small number of unreasonable non-compliers living in the well-ordered society, which, nonetheless, has successfully achieved and maintained stability for the right reasons.

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

So, although Rawls makes it clear that his main focus was in ideal (rather than non-ideal) theory (Rawls 2001: 13), there are grounds to think that Rawls was not completely airtight in his requirement that his well-ordered society must be composed of 100% strict compliers of justice. Rawls did allow a small number of non-compliers to coexist in his well-ordered society. Presumably, this was because he believed that the existence of a small number of non-compliers would not have a devastating effect on a well-ordered society's ability to achieve and maintain stability for the right reasons.

Even without reading too much into this, it is imperative for any theory of political stability to properly discuss how it could cope with the existence or intrusion of non-compliers (if this is not what Rawls had done so already). This is because even a self-contained system like Rawls's well-ordered society can always be infiltrated by a small number of non-compliers either by an external shock or internal mutation or even by a simple mistake made by the reasonable compliers themselves. If the intrusion of a small number of non-compliers disrupts the well-ordered society from properly achieving or maintaining stability for the right reasons, then this would, at the very least, show that, as a theory of political stability, Rawls's stability for the right reasons is not robust.

The purpose of this paper is to examine this particular problem more formally from the perspective of rational choice theory[1] – that is, I will examine the conditions under which full-scale instability emerges from the rational beliefs and preferences of the individuals residing in Rawls's well-ordered society after the intrusion of a small number of non-compliers. In the end, my aim is to show the instability of "stability for the right reasons" in Rawls's well-ordered society. Specifically, I will show how stability for the right reasons can be completely destroyed by the intrusion of an *arbitrarily small number* of unreasonable non-compliers that weakens, ever so slightly, people's common knowledge, which is a defining characteristic of a well-ordered society, which, in turn, is a precondition for stability for the right reasons to be achieved. The analysis that proceeds will be game-theoretic.[2]

1 Here, I am distinguishing rational choice theory from the more empirically based approach that focuses dynamic models of evolution. See Skyrms (2004: 10) and Vanderschraaf (Forthcoming: Introduction) to understand the main difference between the rational choice-based approach and the more empirically based approach. I will later explain in the concluding section why I think the more empirically based approach is inadequate for the purpose of analyzing the particular problem introduced in this paper.

2 There have been many recent papers published in the "public reason liberalism" literature that specifically focus on stability issues of Rawls's well-ordered society. The main question is how people living under a pluralistic liberal democracy, who hold divergent and incompatible moral and religious beliefs, can all affirm the very same political conception of justice regulating the basic structure of their society. Specifically, the problem is this: if I am a reasonable person who upholds and abides by the political conception of justice regulating the basic structure of society, how can I be *assured* that my counterpart would also be the same type of person? Hadfield and Macedo (2012) argue that using *public reason* can play a similar role to that of a 'common logic' in the Hadfield-Weingast positive model (Hadfield and Weingast 2012) and thereby solve this assurance problem. Gaus (2011) and Thrasher and Vallier (2015) argue that public reason will be insufficient to solve the assurance problem as it is mere *cheap talk*. Kogelmann and Stich (2016) argue that we can overcome this too cheap talk problem by resorting to convergence discourse as *costly signals*. We can see that most of the literature in public reason liberalism has focused on using various communicative devices as an assurance mechanism, and, to my best knowledge, no paper has addressed the problem of the common knowledge precondition of Rawls's well-ordered society which this paper tries to address.

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

2. THE MODEL

We wish to present a model of Rawls's well-ordered society. Suppose that our well-ordered society is going through a constitutional change – a change which is practically irrevocable once made. Let N = {1, 2} be the set of two public officials of our well-ordered society, whose mutual agreement would be sufficient to amend the constitution 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 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 free-standing political conception of justice without having the desire to impose her particular comprehensive doctrine to 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 which she believes to be true or 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-interest derived from their own particular conception of the good, and choose the most effective means to achieve it (Rawls 1993/2005: 50).

To simplify our model, we assume that public official 2 is fixed as a reasonable type (i.e. she is type "r".) However, we assume that public official 1 can either be reasonable or unreasonable (i.e. she is 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"). 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: (a) she proposes to amend the constitution with the specific intent to impose her particular comprehensive throughout society, and (b) she disapproves whatever constitutional amendment the other political official proposes. To simplify the model, I will not separately model the proposal and approval stage; proposing a reasonable constitutional amendment and approving the constitutional amendment proposed by the other political official are lumped together under the action label 'cooperate,' while proposing an unreasonable constitutional change and disapproving the constitutional change proposed by the other political official are lumped together under the action label 'defect.' This is not too much loss of generality as 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 non-compliers in terms of their effects on common knowledge.

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 for each:

- *Hegemony* when the political official herself unilaterally defects while the other cooperates;
- *Justice* when both political officials cooperate;
- *Instability* when both political officials defect;

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

- *Subjugation* when the political official herself unilaterally cooperates while 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), while 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 while the other defects (i.e. when she proposes constitutional amendments in accordance with the political conception of justice that gets rejected, and unilaterally approves the proposal made by the other party who intends to impose her particular comprehensive doctrine). The matrix shown in 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 political official's payoffs depend both on the social outcome as well as her type. Let $u_i: O \times T \to \mathbb{R}$ denote player *i*'s payoff function such that

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

where H, J, 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*) players. We can see that while the reasonable types most prefer to achieve mutual justice, the

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

Fig. 1. Social Outcomes in a Well-Ordered Society

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

unreasonable types most prefer to control complete political hegemony, which would allow them to create a perfectionist state in accordance to their moral/religious beliefs. That is, while the reasonable types most prefer to enact laws and implement policies that are contained within the boundaries of a well-ordered liberal democratic society – the characteristic features of which is, what Rawls called, "reasonable pluralism" – the unreasonable types wish to impose their particular comprehensive doctrines to everybody using coercive state power. (It might help to think of the unreasonable types as radical religious fundamentalists.)

The payoffs also show that, regardless of their 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,

> The case for the two principles can be strengthened by spelling out in more detail the notion of a free person. Very roughly the parties regard themselves as having a highest-order interest in how all their other interests, including even their fundamental ones, are shaped and regulated by social institutions. They do not think of themselves as inevitably bound to, or as identical with, the pursuit of any particular complex fundamental interests that they may have at any given time … Rather, free persons conceive of themselves as beings who can revise and alter their final ends and who give first priority to preserving their liberty in these matters. Hence, they not only have final ends that they are in principle free to pursue or to reject, but their original allegiance and continued devotion to these ends are to be formed and affirmed under conditions that are free. (Rawls 1971/1999: 131–2)

In other words, according to Rawls, not only do people have particular interests, but they also have what might be called a *meta*-interest to live under conditions that allow them to freely affirm and revise their particular interests in non-coercive and autonomous ways. 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: $|-S| > J, H > 0$.

We now present our formal definition of stability for the right reasons.

> ***Definition*** (***Stability for the Right Reasons***) A society achieves *Stability for the Right Reasons* if the outcome, *Justice*, can be supported as an equilibrium of the game.

Let G_1 denote the 2×2 game in which political official 2 is reasonable while political official 1 is unreasonable – i.e. G_1 denotes the game in which political official 2 has stag hunt preferences while political official 1 has prisoner's dilemma preferences. Game G_1 is shown in Figure 2.

Many people will recognize that Game G_1 is simply a game that Kavka (1989) called an "assurance dilemma." Game G_1 has a unique Nash equilibrium – viz. *Instability* in which both players defect. Game G_1 represents a situation in which the reasonable people always interact with (and correctly believe that they will always interact with) an unreasonable person. It is not surprising that society is unable to achieve stability for the right reasons in such circumstances.

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

1u \ 2r	c	d
c	$1, J$	$-S, 1$
d	$H, -S$	$0, 0$

Fig. 2. Game G_1 (A game between a unreasonable player 1 and a reasonable player 2)

Let G_2 denote the 2×2 game in which both political officials 1 and 2 are reasonable. (i.e. G_2 denotes the game in which both players 1 and 2 have stag hunt preference.) Game G_2 is shown in Figure 3.

Game G_2 can be seen as a model of a perfectly well-ordered society in which every political official is reasonable – i.e. everybody is a conditional cooperator. One can easily see that game G_2 has two pure strategy Nash equilibria: (a) *Justice* in which both political officials cooperate (i.e. play c), and (b) *Instability* in which both political officials defect (i.e. play d.) Hence, in game G_2, which models a well-ordered society, every political official can affirm (and know that every political affirms) the public conception of justice and achieve stability for the right reasons as a Nash equilibrium.

Let us now examine what may happen after our well-ordered society is introduced with some unreasonable types.

Consider the following game.[3]

Here, there are three possible states, ω_1, ω_2, ω_3, which can each occur with equal probability: $\frac{1}{3}$. The two political officials play the game G_1 in state ω_1, while the two political officials play the game G_2 in states ω_2 and ω_3. Political official 1's information set[4] is outlined by dotted-lines, while political official 2's information set is outlined by dashed-lines. In this game, political official 2, who is herself reasonable, interacts with another reasonable type of political official 1 in every state except in state ω_1, in which case, she interacts with an unreasonable type of political official 1. So, the probability that political official 2 will interact with an unreasonable type of political official 1 is $\frac{1}{3}$. This represents a situation in which a number of unreasonable types have intruded and occupied $\frac{1}{3}$ of the parliament/cabinet of the well-ordered society. Call this game "A Well-Ordered Society after the Intrusion of (1/3) Unreasonable Types."

Suppose that the true state is ω_3. Then, political official 2 knows that the true state is ω_3; so, she knows that she is interacting with another reasonable person. Now, when the true state is ω_3, political official 1 believes that the true state is either ω_2 or ω_3; so, he knows that he is reasonable and also knows that he is interacting with another reasonable person.

3 The game introduced here has a similar logical structure to the game in Osborne (2004: 9.3.2 Infection, 283–5). I will later present and prove a much more general result in this paper.

4 An information set represents what each political official knows about the state of the world in each state. For instance, in state ω_1, political official 1's information set (dotted) signifies that she knows that the true state of the world is ω_1 – that is, she knows that she is an unreasonable type and knows that she is interacting with a reasonable political official. However, in the same state, political official 2's information set (dashed) signifies that she knows that the true state of the world is not ω_3, but she is unsure whether the true state of the world is ω_1 or ω_2 – that is, she knows that she is reasonable, but is unsure whether her counterpart is unreasonable (state ω_1) or reasonable (state ω_2).

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

1r \ 2r	c	d
c	J, J	−S, 1
d	1, −S	0, 0

Fig. 3. Game G_2 (A Game between Two Reasonable Players – i.e. A Well-Ordered Society)

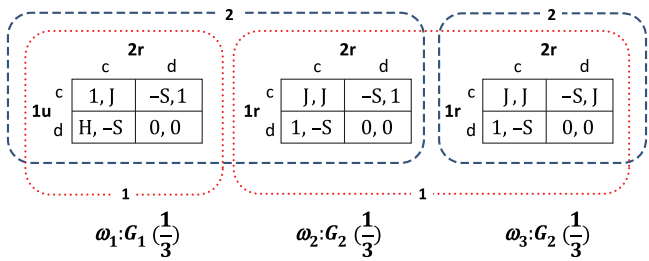


Fig. 4. A Well-Ordered Society after the Intrusion of (1/3) Unreasonable Types

Since political official 2 knows that political official 1 will (when the true state is ω_3) believe that the true state is either ω_2 or ω_3, political official 2 knows that political official 1 knows that political official 2 is a reasonable person. However, when the true state is ω_3, political official 1 will believe that there is one-half chance of ω_2 being the true state, and, if ω_2 is the true state, political official 1 will know that political official 2 will believe that the true state is either ω_1 or ω_2, in which case, political official 2 will believe that there is one-half chance of her interacting with an unreasonable person and a one-half chance of her interacting with a reasonable person. In other words, when the true state is ω_3, political official 1 does not know whether political official 2 knows that political official 1 is a reasonable person; that is, there is uncertainty in political official 1's 2nd-order belief.

At first sight, the situation is favorable for achieving mutual justice since *each political official knows that his/her counterpart is reasonable*. The problem is that one reasonable person (i.e. political official 1) does not know what the other reasonable person (i.e. political official 2) *knows*. Let us see how this can disrupt both reasonable people from achieving stability for the right reasons.

In the game, we would need to specify a strategy for each political official in each of her information sets. Since defecting (i.e. playing d) is a strictly dominant strategy for an unreasonable type, it is optimal for political official 1 (who is unreasonable in state ω_1) to defect (i.e. play d) in state ω_1. Now, consider the information set $\{\omega_1,\omega_2\}$, which belongs to political official 2. Here, political official 2 believes that there is one-half chance that the true state is ω_1, and one-half chance that the true state is ω_2. By playing d, political official 2 can guarantee a payoff of 0. By playing c, political official 2 can receive a payoff of at most[5]: $\frac{1}{2}(-S)+\frac{1}{2}(J)<0$. Therefore, it is optimal for political official 2 to defect in information set $\{\omega_1,\omega_2\}$. A similar reasoning leads political official 1 to defect in information set $\{\omega_2, \omega_3\}$. And, as political official 1 defects in information set $\{\omega_2, \omega_3\}$, player 2 has no

5 This is when 1u defects and 1r cooperates.

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

choice but to defect when the true state is ω_3. Hence, the unique Bayesian Nash equilibrium (BNE)[6] is for both political officials to defect in every information set. This means that our well-ordered society will necessarily destabilize even when the true state is ω_3 in which both parties perfectly know that the other party is reasonable!

What is interesting about this result is that Instability, which is Pareto-dominated by Justice, results not from uncertainty in the other player's *type*, but from uncertainty in the other player's *beliefs*.[7] That is, as political official 2 will defect when she does not know whether political official 1 is reasonable or unreasonable, this forces political official 1 to defect when she does not know whether political official 2 knows that political official 1 is reasonable, and this, in turn, forces political official 2 to defect even when she knows that political official 1 is reasonable and knows that political official 1 knows that political official 2 is reasonable. Such contagion effect can reach further out to even higher-order beliefs.

For instance, consider the following situation in which a *smaller* number of unreasonable types have intruded the political offices of our well-ordered society:

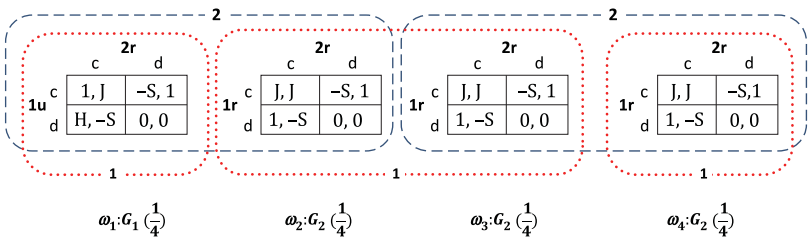


Fig. 5. A Well-Ordered Society after the Intrusion of (1/4) Unreasonable Types

This game is called, "A Well-Ordered Society after the Intrusion of (1/4) Unreasonable Types." Here, the only state in which political official 2, who is herself reasonable, would actually meet a unreasonable person is in state ω_1, which occurs with probability 1/4.

Now, suppose that the true state is ω_4. Then, political official 1 knows that political official 2 is reasonable (i.e. player 1 is certain in his 1st-order belief.) Also, political official 2 knows that political official 1 is reasonable (i.e. political official 2 is certain in her 1st-order belief.) Also, political official 1 knows that political official 2 knows that political official 1 is reasonable (i.e. political official 1 is certain in his 2nd-order belief.) Also, political official 2 knows that political official 1 knows that political official 2 is reasonable (i.e. political official 2 is certain in her 2nd-order belief.) Also, political official 1 knows that political official 2 knows that political official 1 knows that political official 2 is reasonable (i.e. political official 1 is certain in his 3rd-order belief.) However, political official 2 does not know whether political official 1 knows that political official 2 knows

6 Informally, a Bayesian Nash equilibrium (BNE) is a profile of strategies in which every player plays optimally in her information set given the strategies played by other players. See the proof in the Appendix for a more formal definition of a BNE.

7 This is different from recent game-theoretic analysis of Thomas Hobbes's (1994) state of nature in which instability is the result of uncertainty in other people's *types* (viz. whether she is modest or vainglorious) and not in other people's *beliefs*. See Vanderschraaf (2006) and Chung (2015).

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

that political official 1 is reasonable (i.e. there is uncertainty in political official 2's 3rd-order belief.)

It is easy to see that such uncertainty in political official 2's 3rd-order belief triggers a similar contagion effect as before, forcing both players to defect in every information set, which, in turn, results in *Instability* as the game's only equilibrium outcome. Again, the introduction of a smaller number of unreasonable types was still sufficient for society to fail to achieve stability for the right reasons. Such logic allows us to prove the main theorem of this paper.

Theorem: *Let* $\epsilon > 0$ *be the proportion of unreasonable people that have intruded in our well-ordered society. For each* ϵ *arbitrarily close to* 0*, there exists a game of A Well-Ordered Society after the Intrusion of* (ϵ) *Unreasonable Types in which Instability is the unique equilibrium outcome of the game.*

Proof. See Formal Appendix.

Figure 6 shows an example of a game of Well-Ordered Society after the Intrusion of (ϵ) Unreasonable Types (where $\epsilon = \frac{1}{n}$).

Here, the only state in which political official 2, who is reasonable, can ever interact with an unreasonable person is in state ω_1, which can occur with probability $\epsilon > 0$, a number that is arbitrarily close to zero. The situation represents a well-ordered society after the intrusion of an arbitrarily small number of unreasonable types.

Our theorem shows that society will necessarily destabilize even when the true state is ω_n in which case both political officials know that their counterparts are reasonable, both political officials know that their counterparts know that both players are reasonable, both political officials know that their counterparts know that both political officials know that their counterparts are reasonable, and so on until the word "know" occurs n − 2 times. The reason why political official 2 has no choice but to attack in state ω_n is because there is uncertainty in political official 1's $(n-1)^{\text{th}}$-order belief [i.e. political official 1 does not know whether political official 2 knows that political official 1 knows that ... (with the word "know" occurring $n-1$ times) ... that political official 2 knows that political official 1 is reasonable], which renders attacking the optimal strategy for political official 1 in information set $\{\omega_{n-1}, \omega_n\}$, which, in turn, forces political official 2 to attack when she knows that the true state is ω_n.

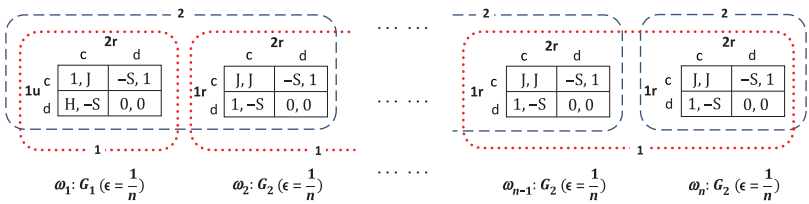


Fig. 6. A Well-Ordered Society after the Intrusion of ($\epsilon = \frac{1}{n}$) Unreasonable Types

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

3. CONCLUSION AND SOME FINAL REMARKS

Our theorem has a very surprising implication for Rawls's notion of stability for the right reasons. Remember that Rawls assumes that in a well-ordered society, "(*nearly*) everyone strictly complies with, and so abides by, the principles of justice" (Rawls 2001: 13, emphasis added). As I have said, the adverb "nearly" allows (or should allow) the intrusion of some unreasonable types in a well-ordered society. Rawls was not exactly clear on how low the proportion of unreasonable types in a well-ordered society has to be kept in order for the well-ordered society to maintain stability for the right reasons. However, no matter how low that threshold is supposed to be, our theorem shows that we could always find a proportion of unreasonable types that is *even lower* that could be *sufficient* to completely destabilize political order in a well-ordered society by undermining common knowledge ever so slightly.

Before ending this paper, I would like to briefly comment on a number of things, which include the generality of the model; the relationship between the specific approach taken in this paper and other approaches used in the literature that focuses on dynamic models of social and cultural evolution; and using costly signals as a potential solution to the problem.

First, the fact that our model consists of only 2 players and 2 types does not imply too much loss of generality. This is similar to the fact that the strategic structures of many collective action problems and public good provision problems can, in many cases, be adequately modeled and analyzed via a simple game such as the 2 player Prisoner's Dilemma (PD game). In such situations, extending the model to an n player PD game does not really change the core structure of the strategic interaction. In both versions (i.e. the 2 player version and the n player version) of the PD game, each individual has a dominant strategy to defect, which renders universal defection as the unique Nash equilibrium, which, in turn, is suboptimal. In these situations, the theoretical gain of extending a 2 player PD game into an n player PD game is not that significant.

The same thing can be said to our model of a well-ordered society after the intrusion of some proportion of unreasonable types. In fact, if one wishes, we could always re-describe the model depicted in Figure 6 (i.e. a well-ordered society after the intrusion of $\epsilon = \frac{1}{n}$ unreasonable types) to include any number of players. The way to do this is to assign each separate information set to a separate player – for instance, we could assign the first information set that contains state ω_1 to player 1, the second information set that contains states ω_1 and ω_2 to player 2, …, and the $(n+1)^{th}$ information set that contains state ω_n to player$n+1$. From this procedure, we can always re-describe the model in Figure 6 as involving *any* number of $n+1$ players, and exactly the same results will follow through. In other words, the central result of this paper does not critically hinge on the fact that our model relied on only 2 players and 2 types.

Of course, all of this depends on our model being a model of a one-shot interaction. I have tried to justify the framework by framing the model to represent a situation in which our well-ordered society is going through a constitutional change – which, once made, is practically irrevocable. However, it would be interesting for future research to see whether the same result will go through if players who have very complicated belief structures as in our one-shot model play the game repeatedly.

Second, note that the problem introduced here cannot be properly identified from the perspective of the more empirically based approach (most notably taken by Skyrms (2004, 2010) and others) that relies on evolutionary game theory under dynamic settings. In

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

evolutionary game theory, each individual in a given population is *not* seen as a *rational agent* – by 'rational agent' I mean an individual who holds beliefs about the world, deliberates, and then chooses the best course of action given his/her beliefs and preferences. Instead, each individual is assumed to play a fixed strategy (say, cooperate or defect). Types are defined by the specific strategy that a given type is fixed to play. In a given round of play, each individual is paired with another individual and each individual simply plays her fixed strategy. Payoffs are defined as the number of offspring the player reproduces for the next round of play. Each offspring plays the same strategy as its parent. Hence, strategies that have done better than average in previous rounds will gradually increase its proportion in subsequent rounds. The main focus of evolutionary game theory is to identify which strategy-types are *evolutionarily stable* – that is, it tries to identify which strategy can successfully repel mutant strategy-types within a given population or successfully infiltrate and take over another population.

The fact that each individual is assumed to play a fixed strategy means that evolutionary game theory does not see individuals as *rational deliberators that can choose*. The fact that types are defined in terms of strategies (and not by each individual's underlying preferences) means that, as long as two individuals perform the same actions, evolutionary game theory is blind to both their cognitive as well as their motivational states. In other words, under evolutionary game theory, a person who defects in order to maximize self-interest and a person who defects because she lacks assurance are epistemically indistinguishable. In our framework, this means that, under evolutionary game theory, neither the distinction between the reasonable and the unreasonable types (whose primary difference lies in each type's underlying preferences) can be properly made, nor can the problem of uncertainty in one's higher-order beliefs, as well as its disruptive effects on stability be properly formulated.

Different tools are apt for analyzing different sets of problems and there are obviously many areas in which the tools of evolutionary game theory shine. However, the problem introduced in this paper, due to its highly epistemic nature, is neither something that can be properly formulated nor solved via dynamic models of evolution. Evolutionary game theory is simply not an adequate tool to analyze the disruptive effects of *uncertainty in higher-order beliefs among rational deliberators*. This is why dynamic models of evolution (à la Skyrms) that reinforce cooperators by ostracizing defectors cannot be directly applied to solve our current problem.

Lastly, one might wonder why our predicament cannot be overcome by the use of *costly signaling*. To pursue this resolution, we must first ask: what theoretical resources does Rawls provide in his political philosophy that could serve the purpose of costly signals? Some have argued that *public reason* can do the trick (see Hadfield and Macedo 2012). The main idea is that by relying solely on public reasons during public deliberation, one can signal that one is a reasonable cooperator who is firmly committed to the political conception of justice, and, thereby, assure other reasonable cooperators to reciprocate cooperation. The problem with this approach is that using public reason is likely to be *cheap talk* (see Gaus 2011; Thrasher and Vallier 2015). That is, due to public reason being 'publicly affordable', an unreasonable type may easily mimic the public signals of the reasonable type during public deliberation for the sheer sake of inducing cooperation from other reasonable types that she could readily exploit.

Others have argued that this gives us reason to move away from the consensus framework of public reason to a more inclusive framework of convergence discourse where people can freely use reasons that stem from their own comprehensive doctrines during public

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

deliberation as long as their views eventually converge toward the public conception of justice (see Gaus 2011; Kogelmann and Stich 2016). Kogelmann and Stich (2016) argue that engaging in convergent discourse can actually serve as a costly signal that could overcome what they call the "too cheap talk problem" of public reason as individuals now have to expend considerable amount of effort to learn other people's comprehensive doctrines to provide reasons that could appeal to other people *qua* endorsers of a particular comprehensive doctrine (Kogelmann and Stich 2016: 724–7). In an as yet unpublished manuscript, Chung (2017) formally demonstrates why Kogelmann and Stich's proposed solution cannot be entirely successful. The point is that even if costly signaling has the potential to solve our current problem, it is not immediately obvious that Rawls has provided the set of theoretical resources that could adequately serve the purpose of costly signals in his framework. In any case, this is an important topic which deserves to be discussed in a separate paper.

To conclude, the main theorem of this paper can be seen as disrupting Rawls's notion of stability for the right reasons in two interrelated ways: (a) by forcing all reasonable types along with all unreasonable types to defect and, thereby, frustrating all possible social cooperation of upholding the same public conception of justice, and (b) by relegating a well-ordered society to a non-well-ordered society by disrupting its common knowledge precondition. The underlying mechanism is that the introduction of an arbitrarily small number of non-compliers causes (b), which, in turn, causes (a), which, in turn, destroys stability for the right reasons. The initial step of this causal chain (viz. the introduction of an arbitrarily small number of non-compliers) is either something that Rawls himself acknowledges or something that Rawls should acknowledge (if he wishes his theory of political stability to be robust), and is something that can happen very easily in all types of political societies (either by external shock or internal mutation or by simple mistake). Hence, our theorem has demonstrated the instability of "stability for the right reasons."

FORMAL APPENDIX – PROOF OF THEOREM

Theorem: *Let* $\epsilon > 0$ *be the proportion of unreasonable people that have intruded in our well-ordered society. For each* ϵ *arbitrarily close to* 0, *there exists a game of A Well-Ordered Society after the Intrusion of* (ϵ) *Unreasonable Types in which Instability is the unique equilibrium outcome of the game.*

Proof. The proof will proceed by construction. Let $\epsilon > 0$ be arbitrary. Define $n = \frac{1}{\epsilon}$ (where x denotes the smallest integer not less than x.) Let $\Omega = \{\omega_1, \ldots, \omega_n\}$be the set of possible states.

Let $G{:}\Omega \to \{G_1, G_2\}$ be the game function that indicates which game is played in each state. Specifically, let $G(\omega_1) = G_1$, and $G(\omega_k) = G_2$ for all $k = 2, \ldots, n$. Let the common prior $p{:}\Omega \to [0, 1]$ for each state occurring be:

$$\begin{cases} p(\omega_k) = \epsilon \text{ for all } k = 1, \ldots, n, \text{ if } \frac{1}{\epsilon} = \frac{1}{\epsilon} = n, \text{ or} \\ p(\omega_k) = \epsilon \text{ for } k = 1, \ldots, n-1 \text{ and } p(\omega_n) = 1 - \epsilon(n-1), \text{ if } \frac{1}{\epsilon} < \frac{1}{\epsilon} = n \end{cases}$$

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

Since public official 2, who is herself reasonable in all states, interacts with a unreasonable person only in state ω_1, which can occur with probability $\epsilon > 0$, this models the fact that the proportion of unreasonable people in society is $\epsilon > 0$.

Let 2^{Ω} denote the set of all subsets of Ω. Then, $\tau_i : \Omega \to 2^{\Omega}$ (for $i = 1, 2$) is public official i's signal function indicating which information set public official i believes to be in given some true state of the world. Specifically, let

$$\begin{array}{ll}
\tau_1(\omega_1) = \{\omega_1\}; & \tau_2(\omega_1) = \tau_2(\omega_2) = \{\omega_1, \omega_2\}, \\
\tau_1(\omega_2) = \tau_1(\omega_3) = \{\omega_2, \omega_3\}; & \tau_2(\omega_3) = \tau_2(\omega_4) = \{\omega_3, \omega_4\}, \\
\vdots & \vdots \\
\tau_1(\omega_{2k}) = \tau_1(\omega_{2k+1}) = \{\omega_{2k}, \omega_{2k+1}\}; & \tau_2(\omega_{2k+1}) = \tau_2(\omega_{2(k+1)}) = \{\omega_{2k+1}, \omega_{2(k+1)}\} \\
\vdots & \vdots \\
\tau_1(\omega_{n-1}) = \tau_1(\omega_n) = \{\omega_{n-1}, \omega_n\}; & \tau_2(\omega_n) = \{\omega_n\} \text{ (when } n \text{ is } odd) \text{ or} \\
\tau_1(\omega_n) = \{\omega_n\}; & \tau_1(\omega_{n-1}) = \tau_1(\omega_n) = \{\omega_{n-1}, \omega_n\} \text{ (when } n \text{ is } even).
\end{array}$$

Let T_i denote the range of τ_i (for $i = 1, 2$). So, T_i is the set of all possible information sets for player i. A strategy for player i is: $s_i : T_i \to \{c, d\}$ – that is, a strategy for public official i assigns an action [either cooperate (c) or defect (d)] for each possible information set in the range of τ_i.

We have just constructed a *Well-Ordered Society after the Intrusion of (ϵ) Unreasonable Types* with n states. The only thing left for us to do is to show that both public officials defecting (i.e. playing d) is the unique equilibrium outcome of the game. Our equilibrium concept is Bayesian Nash Equilibrium (BNE). A BNE in this setting is a profile of strategies (s_1^*, s_2^*) for public officials 1 and 2 such that

$$Eu_i(s_i^*(t_i)|t_i, s_{-i}^*) \geq Eu_i(a_i|t_i, s_{-i}^*)$$

for all i = 1, 2, all $t_i \in T_i$, and all $a_i \in \{c, d\}$. That is, a BNE is a profile of strategies for players 1 and 2 that is optimal for each player in each of their information sets given the strategy played by the other player.

Now, consider the information set $\{\omega_1\}$, which belongs to public official 1. Here, public official 1 knows that she is playing game G_1 – she knows that she is unreasonable. Here, defecting (i.e. playing d) strictly dominates cooperating (i.e. playing c). Therefore, the optimal strategy for player 1 in information set $\{\omega_1\}$ is to defect: $s_1^*(\{\omega_1\}) = d$.

Now, consider the information set $\{\omega_1, \omega_2\}$, which belongs to public official 2. Here, public official 2 believes that the true state is ω_1 with probability $\frac{\epsilon}{\epsilon + \epsilon} = \frac{1}{2}$, and, believes that the true state is ω_2 with probability $\frac{\epsilon}{\epsilon + \epsilon} = \frac{1}{2}$. If the true state is ω_1, then public official 2 knows that she will face a unreasonable person who will defect. So, by defecting (i.e. playing d), public official 2 can guarantee a payoff of at least 0. By cooperating (i.e. playing c), public official 2 can receive at most a payoff of: $\frac{1}{2}(-S) + \frac{1}{2}(J) < 0$. Therefore, the optimal strategy for public official 2 in information set $\{\omega_1, \omega_2\}$ is to defect: $s_2^*(\{\omega_1, \omega_2\}) = d$.

Now, consider the information set $\{\omega_2, \omega_3\}$, which belongs to public official 1. A similar reasoning shows that the optimal strategy for public official 1 in information set $\{\omega_2, \omega_3\}$ is to defect: $s_1^*(\{\omega_2, \omega_3\}) = d$.

Now, assume that it is optimal for public official 2 to defect (i.e. play d) in information set $\{\omega_{2k-1}, \omega_{2k}\}$, and optimal for public official 1 to defect (i.e. play d) in information set $\{\omega_{2k}, \omega_{2k+1}\}$ for some $k \in \mathbb{N}$ such that $2k + 3 < n$.

Now, consider the information set $\{\omega_{2k+1}, \omega_{2k+2}\}$, which belongs to public official 2. Public official 2 believes that the true state is ω_{2k+1} with probability $\frac{\epsilon}{\epsilon + \epsilon} = \frac{1}{2}$, and ω_{2k+2} with probability $\frac{\epsilon}{\epsilon + \epsilon} = \frac{1}{2}$. By assumption, public official 1 will defect (i.e. play d) when the true state is ω_{2k+1}. So, by defecting (i.e. playing d), public official 2 can guarantee a payoff of 0, while by cooperating (i.e. playing c),

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

public official 2 can receive at most a payoff of $\frac{1}{2}(-S)+\frac{1}{2}(J)<0$. Therefore, the optimal strategy for public official 2 in information set $\{\omega_{2k+1}, \omega_{2k+2}\}$ is to defect: $s_2^*(\{\omega_{2k+1}, \omega_{2k+2}\})=d$.

Now, consider the information set $\{\omega_{2k+2}, \omega_{2k+3}\}$, which belongs to public official 1. By a similar argument, the optimal strategy for public official 1 in information set $\{\omega_{2k+2}, \omega_{2k+3}\}$ is to defect: $s_1^*(\{\omega_{2k+2}, \omega_{2k+3}\})=d$.

By mathematical induction, we have shown that for all $k\in\mathbb{N}$ and $2k<n$, the optimal strategy for public official 2 in information set $\{\omega_{2k-1}, \omega_{2k}\}$ is to defect (i.e. $s_2^*(\{\omega_{2k-1}, \omega_{2k}\})=d$), and for all $k\in\mathbb{N}$ and $2k+1<n$, the optimal strategy for player 1 in information set $\{\omega_{2k}, \omega_{2k+1}\}$ is to defect (i.e. $s_1^*(\{\omega_{2k}, \omega_{2k+1}\})=d$.)

Now, consider the state ω_n. Suppose that n is *odd*. Then, $\tau_1(\omega_n)=\{\omega_{n-1}, \omega_n\}$ and $\tau_2(\omega_n)=\{\omega_n\}$.

If $\frac{1}{\epsilon}=\frac{1}{\epsilon}$, public official 1 believes that the true state is ω_{n-1} with probability $\frac{\epsilon}{\epsilon+\epsilon}=\frac{1}{2}$, and ω_n with probability $\frac{\epsilon}{\epsilon+\epsilon}=\frac{1}{2}$. As public official 1 knows that public official 2 will defect (i.e. play d) in information set $\{\omega_{n-2}, \omega_{n-1}\}$, by defecting (i.e. playing d) public official 1 can guarantee a payoff of 0, while by cooperating (i.e. playing c), public official 1 can receive at most a payoff of $\frac{1}{2}(-R)+\frac{1}{2}(P)<0$ in information set $\{\omega_{n-1}, \omega_n\}$. Therefore, the optimal strategy for public official 1 in information set $\{\omega_{n-1}, \omega_n\}$ is to defect: $s_1^*(\{\omega_{n-1}, \omega_n\})=d$. As public official 1 defects (i.e. plays d) in information set $\{\omega_{n-1}, \omega_n\}$, the optimal strategy for player 2 is to defect (i.e. play d) in information set $\{\omega_n\}$: $s_2^*(\{\omega_n\})=d$.

If $\frac{1}{\epsilon}<\frac{1}{\epsilon}=n$, then public official 1 believes that the true state is ω_{n-1} with probability $\frac{\epsilon}{\epsilon+1-\epsilon(n-1)}=\frac{\epsilon}{2\epsilon+(1-\epsilon n)}>\frac{1}{2}$ (as $1-\varepsilon n<0$), and ω_n with probability $1-\frac{\epsilon}{2\epsilon+(1-\epsilon n)}<\frac{1}{2}$. As public official 1 knows that public official 2 will defect (i.e. play d) in information set $\{\omega_{n-2}, \omega_{n-1}\}$, by defecting (i.e. playing d) public official 1 can guarantee a payoff of 0, while by cooperating (i.e. playing c), public official 1 can receive at most a payoff of $\frac{\epsilon}{2\epsilon+(1-\epsilon n)}(-S)+\left(1-\frac{\epsilon}{2\epsilon+(1-\epsilon n)}\right)(J)<\frac{1}{2}(-S)+\frac{1}{2}(J)<0$ in information set $\{\omega_{n-1}, \omega_n\}$. Therefore, the optimal strategy for public official 1 in information set $\{\omega_{n-1}, \omega_n\}$ is to defect: $s_1^*(\{\omega_{n-1}, \omega_n\})=d$. As public official 1 defects (i.e. plays d) in information set $\{\omega_{n-1}, \omega_n\}$, the optimal strategy for public official 2 is to defect (i.e. play d) in information set $\{\omega_n\}$: $s_2^*(\{\omega_n\})=d$.

Now, suppose n is *even*. Then, $\tau_2(\omega_n)=\{\omega_{n-1}, \omega_n\}$ and $\tau_1(\omega_n)=\{\omega_n\}$.The remainder of the proof proceeds in a similar fashion with the roles of players 1 and 2 switched.

We conclude that the unique BNE of this game is for public officials 1 and 2 to defect (i.e. play d) in every information set, and, hence, the unique equilibrium outcome is Instability. ∎[8]

REFERENCES

Chung, H. 2015. 'Hobbes' State of Nature: A Modern Bayesian Game-Theoretic Analysis.' *Journal of the American Philosophical Association*, 1(3): 485–508.

—— 2017. 'Public Reason, Convergence Discourse, and the Well-Ordered Society under Crisis.' Unpublished manuscript, https://ssrn.com/abstract=2862832.

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

—— 2016. 'Consensus on What? Convergence for What?: Four Models of Political Liberalism.' Unpublished manuscript.

8 I thank Jerry Gaus, Brian Kogelmann for helpful suggestions on earlier versions of this paper. I thank the editor and an anonymous reviewer for their constructive criticisms and helpful comments.

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press

Hadfield, G. and Macedo, S. 2012. 'Rational Reasonableness: Toward a Positive Theory of Public Reason.' *Law and Ethics of Human Rights*, 6(1): 1–46.
—— **and Weingast, B.** 2012. 'What is Law? A Coordination Model of the Characteristics of Legal Order.' *Journal of Legal Analysis*, 4(2): 471–514.
Hobbes, T. 1994. *Leviathan* (with selected variants from the Latin edition of 1668). Edited, with Introduction, by E. Curley. Indianapolis, IN: Hackett.
Kavka, G. 1989. 'Political Contractarianism.' Unpublished manuscript.
Kogelmann, B. and Stich, S. 2016. 'When Public Reason Fails Us: Convergence Discourse as Blood Oath.' *American Political Science Review*, 110(4): 717–30.
Lewis, D. 2002. *Convention – A Philosophical Study*. Oxford: Blackwell.
Osborne, M. 2004. *An Introduction to Game Theory*. Oxford: Oxford University Press.
Rawls, J. 1971/1999. *A Theory of Justice*, Revised edition. Cambridge, MA: Harvard University Press.
—— 1993/2005. *Political Liberalism – Expanded Edition*. New York, NY: Columbia University Press.
—— 2001. *Justice as Fairness – A Restatement*. Cambridge, MA: Belknap Press of Harvard University Press.
Skyrms, B. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press
—— 2010. *Signals – Evolution, Learning, and Information*. Oxford: Oxford University Press.
Thrasher, J. and Vallier, K. 2015. 'The Fragility of Consensus: Public Reason, Diversity and Stability.' *European Journal of Philosophy*, 23(4): 933–54. doi: 10.1111/ejop.12020.
Vanderschraaf, P. 2006. 'War or Peace?: A Dynamical Analysis of Anarchy.' *Economics and Philosophy*, 22: 243–79.
—— Forthcoming. 'Learning Bargaining Conventions.' *Social Philosophy and Policy*.

HUN CHUNG has received his PhD in Philosophy at Cornell University. He is currently a PhD candidate in Political Science at the University of Rochester. Previously, he was a Postdoctoral Fellow in PPE and Philosophy at the University of Arizona, and a Visiting Assistant Professor in Philosophy at the Rochester Institute of Technology. His main areas of research interest are in PPE (Philosophy, Politics, and Economics), Ethics/Political Philosophy, and Formal Theory (viz. Decision/Game/Social Choice Theory) (Homepage: http://hunchung.com).

https://doi.org/10.1017/epi.2017.14 Published online by Cambridge University Press