Public Reason's Chaos Theorem[1]

1. Introduction

This paper examines to what extent public reason can serve as an assurance mechanism as, I argue, John Rawls intended it to. A note on terminology: by "public reason" I do not refer to a strand of liberal theories all holding that, understood broadly enough, authoritative commands of a certain type must be justified to all suitably idealized persons in society. By "public reason" I instead refer to the narrower class of norms regulating public discourse in democratic societies: without getting too technical too quickly, public reason requires citizens justify certain classes of authoritative commands to one another using reasons all endorse. Though related, these two uses of "public reason" are distinct (Vallier 2015). As such, this paper examines to what extent public reason – understood as a set of norms regulating democratic discourse – is capable of serving as an assurance mechanism. My thesis is that public reason cannot serve this function.[2]

The structure of this paper is as follows. In the next section I articulate the assurance problem Rawls thought the well-ordered society faced, and how he thought public reason solved this problem. In §3 I link up the literature on public reason's incompleteness to public reason's ability to serve as an assurance mechanism. Namely, I argue that the kind of incompleteness most Rawlsians grant is ubiquitous but unproblematic from a normative standpoint is devastating from an assurance perspective: this form of incompleteness makes it possible for citizens to argue for policy

[1] The author would like to thank Brian Albrecht, Hun Chung, Jerry Gaus, Tim Kearl, and Stephen G.W. Stich for comments on earlier drafts of this paper, as well as Sameer Bajaj for helpful discussion.

[2] I thus join a growing literature that casts doubt on this claim, though the argument given here is unique and has yet to be articulated. See Gaus (2011); Thrasher and Vallier (2015); Kogelmann and Stich (2016); Chung (forthcoming).

conclusions that are favored by their private interests, rather than justice. In response, perhaps the thing to do is structure deliberative democratic institutions such that citizens will always be incentivized to use public reasons to only argue for conclusions they believe are favored by justice. §4 proves that this is impossible by extending the Gibbard-Sattherthwaite theorem. §5 entertains an objection to this paper's main argument, by refining the idea of public reason. It is shown that such refinement fails.

2. Assurance and Stability

It was not enough for Rawls to offer a model of the just society. Rawls also felt incumbent to show that such a society would remain just. Indeed, ensuring that a society regulated by justice as fairness would be stable was integral in justifying the conception of justice in the first place. Upon a close reading, Rawls seems to think there are two kinds of instabilities just societies face. He illustrates both in this passage:

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

The first kind of instability Rawls refers to is a Prisoner's Dilemma, illustrated in Figure 1. In this game, Althea most prefers to act unjustly while Bertha acts justly and vice versa; Althea and Bertha next most prefer mutual adherence to justice; the third most preferred outcome is where both act unjustly; and finally, Althea and Bertha's least preferred outcome is where they act justly and the opposing player acts unjustly, making them a sucker. The only equilibrium in this game is where

both players act unjustly, achieving the Pareto inefficient (2, 2) payoff. Absent intervention, rational players playing Prisoner's Dilemmas with one another will not do as justice requires.

		Bertha	
		Act Justly	Act Unjustly
Althea	Act Justly	3, 3	1, 4
	Act Unjustly	4, 1	2, 2

Figure 1

The second kind of instability Rawls refers to is an assurance dilemma, illustrated in Figure 2. In this game, both Althea and Bertha most prefer to act justly, which already differs from the Prisoner's Dilemma. Althea's second most preferred outcome is where she acts unjustly and Bertha does not, and vice versa; the third most preferred outcome for both players is where both act unjustly; and each players' least preferred outcome is where she acts justly while the other player acts unjustly, once again making them a sucker. In assurance games there are two equilibria: a Pareto efficient equilibrium where both Althea and Bertha act justly, achieving the (3, 3) payoff, and a risk-dominant equilibrium where both Althea and Bertha act unjustly, achieving the (1, 1) payoff. Even though acting justly is an equilibrium strategy in this game, and even though mutual adherence to justice yields the efficient equilibrium, it is by no means obvious that rational players will play such a strategy: the slightest doubt that Bertha will act justly can force Althea to act unjustly so as to avoid the sucker payoff, where she acts justly and Bertha acts unjustly (Aumann 2000). Brian Skyrms nicely summarizes the problem: "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).

		Bertha	
		Act Justly	Act Unjustly
Althea	Act Justly	3, 3	0, 2
	Act Unjustly	2, 0	1, 1

Figure 2

On my reading of Rawls, the first threat of instability (the Prisoner's Dilemma) occurs first, followed by the second threat of instability (the assurance problem) once the first threat of instability has been removed. To solve the Prisoner's Dilemma, something about the game must change because the only equilibrium solution is unjust behavior: we need to alter the game in some way so that just behavior is now an equilibrium strategy. Rawls accomplishes this within his framework by arguing that citizens in the well-ordered society develop a sense of justice via a three-stage developmental process (Rawls 1971: ch. 8). Once the sense of justice has been inculcated there is a "desire to do what is just," such that "no one wishes to advance his interests unfairly to the disadvantage of others; this removes instability of the first kind" (Rawls 1971: 497).

But we are not out of the woods yet. As Rawls notes, "even with a sense of justice men's compliance with a cooperative venture is predicated on the belief that others will do their part; citizens may be tempted to avoid making a contribution when they believe, or with reason suspect, that others are not making theirs" (Rawls 1971: 336). That is, even after we have extricated ourselves from the Prisoner's Dilemma we still face the assurance problem. Though citizens want to do what

is just due to their sense of justice, they do not want to do what is just at any cost. If Bertha does not act justly then Althea does not want to act justly either; and if Althea does not act justly then Bertha most prefers to act unjustly as well. To solve this problem, Althea and Bertha need some way of *assuring* one another that they will play the act justly equilibrium, rather than play the act unjustly equilibrium.

On one reading of Rawls's later thought, the assurance problem is solved by public reason (Weithman 2010: 327; Weithman 2015; Hadfield and Macedo 2012). This might not be obvious, because public reason seems to have a normative purpose. The liberal principle of legitimacy, Rawls tells us, requires we exercise political power in a manner justifiable to all. This creates a "moral, not a legal, duty" – the duty of civility – which requires citizens "be able to explain to one another… how the principles and policies they advocate and vote for can be supported by the political values of public reason" (Rawls 1993/2005: 217). This normative function aside, public reason also serves a practical role: when Althea appeals to strictly public reasons in her discourse rather than reasons taken from her comprehensive doctrine, she signals commitment to justice over her private interests. This assures Bertha that Althea will act justly. And when Bertha uses strictly public reasons in her discourse with Althea rather than reasons taken from her comprehensive doctrine, she assures Althea that she will act justly as well by signaling her commitment to the political conception of justice over her own private interests. On Rawls's model, then, citizens achieve the efficient (3, 3) equilibrium rather than the risk-dominant (1, 1) equilibrium because citizens adhere to the ideal of public reason when engaged in democratic discourse. When this occurs, the second kind of instability is removed.

3. Two Kinds of Incompleteness, Two Kinds of Assurance Breakdowns

We now examine the relationship between the completeness of public reason and public reason's ability to serve as an assurance mechanism. By completeness, I mean public reason's ability to give an answer to all questions to which it is supposed to apply. If public reason is incomplete then it does not provide citizens the means to answer some questions that citizens are supposed to answer with public reasons alone. Though there is a large literature examining to what extent (*i*) the incompleteness of public reason likely obtains, and (*ii*) how damning incompleteness really is when it does obtain, no one has yet examined the relationship between incompleteness and assurance. The purpose of this section is to argue that the one kind of incompleteness those in the literature (*iii*) agree will occur quite frequently, but (*iv*) do not think is normatively problematic, is *incredibly* problematic from an assurance perspective. Granting (*iii*), this casts serious doubt on public reason's ability to serve as an assurance mechanism.

3.1 *Indeterminacy*.

Following Schwartzman (2004) who follows Gaus (1996), we can distinguish between two different ways public reason can be incomplete. The first kind of incompleteness is called *indeterminacy*. Public reason is indeterminate if citizens do not have sufficient reasons to adjudicate between different options they are confronted with in the public sphere. More formally, suppose r_1, r_2, r_3,…, r_n are the public reasons available to citizens. Suppose citizens face policy options p_1 and p_2. Public reason is indeterminate if, taking all r_1-r_n into consideration, citizens cannot muster a sufficiently rigorous argument for either p_1 or p_2.

Here's an example of public reason's indeterminacy. Suppose, following some, that public reason should apply to *all* political questions where citizens exercise coercive power over one another (Quong 2004). Here is one such question that must be answered: what should the monetary policy of the Federal Reserve be? Ought the Fed adopt some form of quantitative easing, or ought they stick with strict monetarism? Since public reasons include positive as well as normative considerations, and since these positive considerations must be "found in common sense," must be "not controversial," and must be "plain truths now widely accepted, or available, to citizens generally," it is doubtful that citizens can answer this particular policy question with public reasons alone (Rawls 1993/2005: 224-225). Given this indeterminacy, citizens will need to appeal to considerations that are not public reasons to adjudicate the relevant dispute. For this particular question, public reason is indeterminate.

Because citizens must appeal to non-public reasons in adjudicating those matters for which indeterminacy obtains it follows that there will be some questions where citizens will not give other citizens reasons that are justifiable to them and, more relevant to our concerns, there will be some questions where citizens will not be able to assure their fellow citizens that they remain faithful to the political conception of justice. If Althea gives an argument for p_1 over p_2, and must appeal to non-public reason r` from her comprehensive doctrine because public reasons $r_1$-$r_n$ are insufficient to justify either $p_1$ or $p_2$, then Bertha might interpret Althea's appeal to r` as signaling fidelity to Althea's private interests over the political conception of justice. Given this signal, Bertha, wanting to avoid the sucker payoff from Figure 2, might end up acting unjustly in anticipation of Althea acting unjustly. Assurance has thus broken down. Because of public reason's indeterminacy, there will be times when Althea and Bertha simply cannot appeal to public reasons; since public reasons are meant to assure other citizens, it follows that there will be times when Althea and Bertha cannot assure one another.

Now whether incompleteness in the form of indeterminacy is a big problem for public reason as an assurance mechanism depends on how often indeterminacy obtains: given that Althea and Bertha engage in repeated interactions, Althea's not adhering to public reason in *one* interaction would not be sufficient to spook Bertha into acting unjustly if Althea adheres to public reason in most other interactions. But, if indeterminacy is the rule and not the exception, then indeterminacy undermines public reason's assurance function: if, for the majority of political questions, there are insufficient public reasons for adjudicating the matter at hand, then assurance will likely break down, leading us to the inefficient equilibrium where everyone acts unjustly.

Hence the fundamental question: how likely is it that incompleteness in the form of indeterminacy obtains? Fortunately, many in the literature think that indeterminacy will be a rather rare problem for public reason. Schwartzman (2004: 205-208) offers compelling argument as to why indeterminacy will be the exception rather than the rule. Here we will grant that Schwartzman is correct; doing so entails that incompleteness in the form of indeterminacy is not a problem for public reason as an assurance mechanism. The reason we do this is because we do not want the critique of public reason as an assurance mechanism to rest on controversial speculation concerning how often certain phenomena occur. Thus, we grant the defender of public reason everything they grant themselves. In the next section we examine a kind of incompleteness that most hold is ubiquitous.

3.2 *Inconclusiveness.*

The second kind of incompleteness is *inconclusiveness*. Public reason is inconclusive if there are competing public reason-based justifications for a given political question that do not dominate one another. More formally, suppose r_1, r_2, r_3,…, r_n are the public reasons available to citizens. Suppose

citizens face policy options p_1 and p_2. Public reason is inconclusive if, taking all r_1-r_n into consideration, citizens can muster sufficiently rigorous arguments for both p_1 and p_2, such that we cannot judge the argument for p_1 to be better than the argument for p_2 or vice versa. Unlike indeterminacy, inconclusiveness *does* allow us to actually argue for different policy options using strictly public reasons; the problem now, though, is that public reasons do not end the argument. Public reasons can be called upon to muster support for both p_1 and p_2 – in such a case, public reason alone does not settle the matter at hand.

Unlike indeterminacy, it is not clear that inconclusiveness ends up requiring citizens appeal to non-public reasons. Instead, after citizens give public reasons in defense of their preferred policies – and, by hypothesis, the matter has yet to be solve – there are other ways of adjudicating the disagreement: we can just hold a simple majority vote between p_1 and p_2, for instance. *Contra* cases of indeterminacy, in cases of inconclusiveness political matters can be settled without appealing to non-public reasons (Schwartzman 2004: §3). Since Rawls's general worry is that the use of non-public reasons signals fidelity to private interests over the political conception of justice, and since cases of inconclusiveness do not require citizens appeal to non-public reasons, it might be thought that inconclusiveness – unlike indeterminacy – does not threaten public reason's ability to function as an assurance mechanism.

This is incorrect. The worry with inconclusiveness is that citizens might not be using public reasons to argue for conclusions they believe to be favored by justice, but rather are arguing for conclusions that are favored by their private interests. For example, suppose Althea thinks one of the policy options (either p_1 or p_2) is favored by the political conception of justice, but the other policy option *not* favored by justice is favored by her comprehensive doctrine. Assuming that public reason is inconclusive in this case, public reasons r_1-r_n support both p_1 and p_2. The worry is that,

when Althea argues for either p_1 or p_2 with strictly public reasons, she might be merely trying to advance her own private interests rather than the demands of justice. If Bertha suspects as much then she might try to be the first to act unjustly in order to avoid the sucker payoff – if Althea is not to be guided by justice in her public deliberators then she does not wish to be either. Note, this problem does not arise if inconclusiveness does not obtain. If the public reasons available entail one and only conclusion then Bertha cannot doubt the conclusion Althea reaches with public reasons. It is only when public reason supports multiple conclusions that we are led to an assurance breakdown.[3]

Like before, inconclusiveness is only a problem if it occurs frequently. If, most the time, public reason entails one and only one conclusion, then Althea using public reason in cases of inconclusiveness every now and then will not cause an assurance breakdown. Unlike indeterminacy, though, it seems likely that inconclusiveness will be quite prevalent. The reason why is that inconclusiveness seems to be a function of how we *weigh* or *tradeoff* different values, and the assignment of different weights and tradeoffs is to be expected – by Rawls's own admission – in liberal societies. Though we all agree that reasons r_1-r_n are the relevant considerations, if Althea thinks r_2 to be more weighty than r_1, and yet Bertha thinks r_1 to be more weighty than r_2, then they will likely be led to different policy conclusions. But note, Rawls tells us that we should *expect*

[3] It might be thought that inconclusiveness between only two policy options cannot cause an assurance breakdown: after all, what are the chances that either p_1 or p_2 line up with Althea's private interests, thus allowing her to argue for her private interests using public reasons? It might be that assurance breakdowns are caused by inconclusiveness only when the set of policies public reason is inconclusive over is quite large, making it more likely that one of the policy options falls within the purview of Althea's private interests. Note, though, that we should *expect* the set of policy conclusions public reason is inconclusive over to be quite large. Below it is argued that, *at the very least*, each unique ranking of the importance of the different reasons in the set of public reasons implies a unique policy conclusion. Since the content of public reason will be quite large (Rawls 1993/2005: 223-227), this implies a good deal many permutations on rankings, and thus a good deal many policy conclusions public reason can be inconclusive over, making it likely that one of the policies in the inconclusive set is favored by Althea's private interests. This is then sufficient for assurance to break down.

disagreement over how to weigh and tradeoff different values. One of the "burdens of judgment," according to Rawls, is that "even where we agree fully about the kinds of considerations that are relevant, we may disagree about their weight, and so arrive at different judgments" (Rawls 1993/2005: 56). Since (*i*) differences in tradeoffs are what make inconclusiveness obtain, and since (*ii*) Rawls tells us that differences in tradeoffs are part of the normal circumstances of everyday human life in liberal societies, we should expect inconclusiveness to be widespread. But this just means public reason fails as an assurance mechanism.

The above relationship concerning tradeoffs and inconclusiveness might have been a bit abstract, so let's consider a concrete example. In an often-discussed footnote of *Political Liberalism* Rawls says that "the troubled question of abortion," involves "three important political values: the due respect for human life, the ordered reproduction of political society over time, including the family in some form, and finally the equality of women as equal citizens" (Rawls 1993/2005: 243n). Though Rawls himself held that "any reasonable balance of these three values" will grant women the right to terminate pregnancies, many have objected that this is not so (de Marneffe 1994: 234; Quinn 1997: 150). Indeed, it seems rather straightforward that different ways of ranking these same three values will lead to radically different policy conclusions. With these three values, there are six logically possible ways of ranking their importance, illustrated in Table 1.

Althea	**Bertha**	**Cassidy**	**Dupree**	**Esau**	**Franklin**
Respect for Human Life	Respect for Human Life	Ordered Reproduction	Ordered Reproduction	Equality of Women as Equal Citizens	Equality of Women as Equal Citizens
Ordered Reproduction	Equality of Women as Equal Citizens	Respect for Human Life	Equality of Women as Equal Citizens	Respect for Human Life	Ordered Reproduction
Equality of Women as Equal Citizens	Ordered Reproduction	Equality of Women as Equal Citizens	Respect for Human Life	Ordered Reproduction	Respect for Human Life

Table 1

Here, because Althea takes respect for human life to be most important followed by the ordered reproduction of society, it is likely Althea can argue for rather stringent restrictions – perhaps complete prohibition – on abortion with public reasons alone. Bertha, who also holds respect to be most important but thinks equality of women is quite weighty as well, would likely also justify restrictions on abortion, but of a less stringent kind: perhaps abortions can be carried out under heavily regulated conditions. Because Cassidy thinks ordered reproduction is most important followed by respect for human life, she could likely endorse restrictions on abortion that serve society's interest in rearing the next generation: something like a quota on abortions, perhaps. Dupree similarly ranks ordered reproduction highly but thinks equality of women as citizens is important: from his perspective the optimal policy might be some kind of minimal regulation of abortion that still grants great autonomy of choice to women. Esau and Franklin, because they rank equality of women as citizens highest, will likely place no or very limited restrictions on abortion: but because Esau takes respect for human life to be second in importance, and because Franklin takes ordered reproduction to be second in importance, they will likely disagree over the nature of these minimal restrictions.

Though the inconclusiveness of public reason seems prevalent, many in the literature do not consider this to be a major worry (Freeman 2007: 242-243; Quong 2011: 209; Quong 2014: 267; Rawls 1993/2005: 240–241; Schwartzman 2004; Williams 2000). Such dismissals of inconclusiveness, though, take as their focus the normative function of public reason: even though we disagree over the relevant policy conclusions, so long as we give each other the same reasons we all share, we treat each other in a manner justifiable to all. Inconclusiveness thus does not prevent public reason from serving its normative function. This paper, though, focuses on public reason as

an assurance mechanism: insofar as we care about this function of public reason, inconclusiveness is a rather deep worry.

Let us take stock. We have seen (*i*) that different weightings of values leads to inconclusiveness, (*ii*) that we should expect different weightings of values to be ubiquitous in liberal societies, and (*iii*) that any time there is inconclusiveness there is the potential worry that citizens advance arguments with public reasons in order to support their private interests rather than what they think justice requires. Since (*i*) and (*ii*) imply inconclusiveness will be prevalent, because of (*iii*) we should seriously doubt public reason's ability to serve as an assurance mechanism. Since inconclusiveness is something we seem to be stuck with, perhaps the best response here is to in some way structure democratic deliberation such that people must *always* give reasons in the order they think best reflect the requirements of justice, and thus *always* support policy conclusions they think are favored by justice. If such were the case, then although inconclusiveness obtains, Althea and Bertha would still assure one another when they give each other public reasons, because they by hypothesis know the other person advances what she thinks justice requires rather than their private interests *even though* what justice requires is inconclusive. We examine this proposal in the next section.

4. Strategy-Proof Democratic Deliberation

The proposal on the table is that we somehow structure democratic deliberation such that Althea and Bertha must always give public reasons in the order they think is favored by justice, and thus must always support those policy conclusions they think are favored by justice rather than their

private interests.[4] If such were the case, then public reason can still serve as an assurance mechanism even when it's inconclusive: institutions guarantee that Althea is not using public reasons to cleverly argue for her own private interests over what justice requires. This assures Bertha. This section proves that this proposal is logically impossible: no institutional arrangement satisfying minimal requirements of democratic equality is capable of achieving the stated goal. To show this we appeal to, and extend, seminal results in social choice theory. Before doing so, though, we take a brief excursion concerning the relevance of social choice theory to political philosophy.

4.1 *Social Choice Theory and Value Tradeoffs.*

Inaugurated by Kenneth J. Arrow's impossibility theorem, social choice theory is usually defined as the study of the mathematical properties of preference aggregation. Typically, this is understood most concretely in terms of voting systems: social choice theory shows how majority rule is almost never in equilibrium, how all voting systems can be manipulated, how different voting rules applied to the same polity can yield radically different electoral results, etc.

This is an unduly narrow interpretation of the social choice theoretic framework, and one that has been pushed back against in recent scholarship. Instead of understanding social choice theory as delivering interesting results about the aggregation of preferences or the properties of voting systems, we should instead understand social choice theory as casting insight on decision-making (both individual and group) whenever we face tradeoffs in the criteria we use to make decisions – which, of course, is almost always. More generally, we can interpret social choice theory's

[4] The notion of structuring democratic deliberation to circumnavigate common problems that plague group discourse (group polarization, opinion leaders, etc.) is frequently appealed to in the deliberative democracy literature. For just one example of how one might actually do this, see Sunstein and Hastie (2014: ch. 6).

focus on "aggregation" to "refer to any process by which two or more goals or criteria are combined or contrasted to yield a final evaluation, prediction, or prescription" (Patty and Penn 2015: 52). As examples of what sorts of things a wider-ranging interpretation of social choice theory can be applied to, social choice has recently been used in the analysis of theory choice in the philosophy of science (Okasha 2011; Stegenga 2015; Morreau 215), paradoxes in decision theory (Briggs 2010), as well as the problem of what to do in the face of "normative uncertainty" (MacAskill 2016).

On this broader understanding of social choice theory, we can understand "preferences" as just different ways of ranking the importance of competing values that go into making a decision: for instance, that equality of women as equal citizens is a more weighty consideration than the protection of human life, which is a more weighty consideration than the ordered reproduction of society. Moreover, we can understand social choice axioms – Universal Domain, Pareto, Non-Dictatorship, etc. – more broadly as well. Instead of merely understanding these rules as formal properties of voting systems, we can understand them as describing processes that are not strictly speaking formal voting procedures: for example, what kinds of formal properties characterize deliberative democratic processes from a descriptive point of view? Or, from the standpoint of a normative ideal, what kinds of formal properties *should* characterize deliberative democratic systems? After answering these questions, we can gain better insight, via the results of social theory, concerning the tenability of the current proposal.

4.2 *The Manipulability of Public Reason*.

Given this broader understanding of social choice theory, we proceed as follows. First, we construct a basic model to better understand the relationship between value tradeoffs and policy conclusions. Then, we propose some minimal characteristics that would ideally characterize *any*

deliberative democratic process. From there we ask whether such an institutional arrangement is *strategy-proof*: that is, will *any* institutions satisfying the proposed criteria always incentivize citizens to elicit their true rankings of public reasons in terms of what they think justice requires? It is then shown that there exists no set of institutions that can ensure this. The current proposal thus fails, leaving us in doubt about public reason's ability to serve as an assurance mechanism.

Inconclusiveness obtains because there are many different values included in the set of public reasons, and these values can be weighed and traded-off in different ways. Above in Table 1 we noted three such values – respect for human life, the ordered reproduction of society, and equality of women as equal citizens. To simplify things, let us suppose that each value in the set of public reasons is represented by a variable. So, x represents respect for human life, y represents the ordered reproduction of society, and z represents equality of women as equal citizens, and so on and so forth if there are more values present. This transforms Table 1 into Table 2, illustrated below.

Althea	**Bertha**	**Cassidy**	**Dupree**	**Esau**	**Franklin**
x	x	y	y	z	z
y	z	x	z	x	y
z	y	z	x	y	x

Table 2

Above it was argued that different ways of weighing and trading-off these values results in different policy conclusions. Simplifying things, we will suppose that each unique way of ordering the values that make up the set of public reasons imply one and only one policy conclusion. Sticking with the simple model illustrated in Table 2, and letting "$\succ$" represent "is ranked above," {x $\succ$ y $\succ$

z} implies policy p_1, {x ≻ z ≻ y} implies p_2, {y ≻ x ≻ z} implies p_3, etc.[5] Note, it is likely that one ordering of values can imply more than one unique policy conclusion; this makes things even worse for the Rawlsian. For simplicity we stipulate that each unique ordering of the values that make up public reason imply one and only one policy conclusion.

We now introduce the notion of a *social welfare function*. A social welfare function is defined as a rule or process that takes every individual's ordering of the relevant values – Althea's {x ≻ y ≻ z}, Bertha's {x ≻ z ≻ y}, etc. – and, from there, constructs one social ordering of the relevant values – say, {y ≻ z ≻ x}. Usually, social welfare functions are understood in the context of voting rules: each individual submits a ballot ranking the candidates and, after application of the relevant electoral rules, one social ranking of the candidates emerges. But, following the analysis in §4.1, we can understand less concrete practices as embodiments of social welfare functions. For our purposes, we can interpret deliberative democratic procedures in such a way. Each participant enters the public sphere with a ranking of values. They engage in reasoned discourse with one another, construct arguments, and hurtle objections followed by rejoinders: at the end of the process, there is a clear hierarchy of values that represent the results of the deliberative process. As an example, after Althea, Bertha, and all the rest end their deliberation, perhaps ordered reproduction of society emerges as most important, followed by equality of women as equal citizens, with respect for human life taking up the rear, and thus {y ≻ z ≻ x}. From here the policy society ought to implement is the unique one attached to {y ≻ z ≻ x}, which is p_3.

Beyond characterizing deliberative democratic institutions as social welfare functions, we can inquire about what sorts of characteristics such institutions ideally would have. For instance, most

[5] Note, we can also allow for the possibility of indifference, so that {x ≻ y ≈ z} implies a different policy than {x ≻ y ≻ z}, where "y ≈ z" represents something like "neither z is ranked above y nor y is ranked above z."

hold it undesirable if deliberative democracy results in polarization of viewpoints; we can stipulate that, ideally, polarization would not happen. Three necessary (but not sufficient) conditions any set of deliberative democratic institutions ought to satisfy are as follows:

> *Pareto*: if *every* individual thinks one value is more important than another value – say, equality of women as equal citizens is more important than respect for human life – then the social ordering of these values must reflect this: that is, equality of women must be socially ranked higher than respect for human life.
>
> *Universal Domain*: all logically possible ways of ordering the different values are permissible.
>
> *Non-Dictatorship*: there is no one individual whose ordering of the values *always* determines the social ordering: that is, the social ordering cannot mimic whatever Althea's ordering is each and every time.

These all seem to be minimal conditions that any ideal of deliberative democracy ought to satisfy. Pareto is straightforward: if everyone thinks that x is more important than y, then x ought to be socially ranked above y when the dust of deliberation settles. Universal Domain says people can tradeoff values in any way they see fit – this seems to follow from Rawls's admission that different weightings are part of the burdens of judgment. Note, Universal Domain does *not* say people may bring in any values they wish into public discourse – it does not force us to take seriously Nazis or racists or anyone else we deem unreasonable. Instead, Universal Domain says that once we have decided on the set of public reasons, *then* citizens can weigh and tradeoff these values in any way they see fit.[6] And finally, Non-Dictatorship simply says that, ideally, one person does not run the whole show. Certainly this follows from the ideal of democratic equality. Though we may be plagued

[6] Plausibly, the most reasonable axiom to relax here is Universal Domain, and there are passages from Rawls's text suggesting that this might be the best route forward (Rawls 1993/2005: 243n). Domain restrictions will not do much work in preventing manipulability, though. As Penn et al. (2011) show, traditional domain restrictions such as single-peakedness that are sufficient to prevent Arrovian cycles are *not* sufficient to prevent manipulability. The kind of domain restriction required to block the theorem will likely be quite extreme indeed, and thus be at odds with Rawls's postulate of reasonable pluralism.

by "opinion leaders" in the real world, recall that we are here characterizing the features of a deliberative ideal.

Having constructed the basic model, we can now explore the proposal. The proposal is that, in order to skirt around the assurance breakdown caused by public reason's pervasive inconclusiveness, we should somehow structure deliberative democratic institutions such that citizens *always* have an incentive to reveal their justice-orderings rather than fake orderings they may take on to argue for conclusions favored by their private interests. A social welfare function is *strategy-proof* if it can do just that. If we can design deliberative democratic institutions such that they are strategy-proof then inconclusiveness does not cause an assurance breakdown: even though public reason supports many different policy conclusions, Bertha by hypothesis knows that Althea, when she gives public reasons, gives them in the order she believes is favored by justice. She thus has no reason to doubt that Althea remains faithful to the political conception of justice in cases of inconclusiveness. Unfortunately, it can be shown that it is impossible to design deliberative institutions so that they are strategy-proof while also satisfying those minimal criteria outlined above.

> **THEOREM:** There is no social welfare function satisfying Pareto, Universal Domain, and Non-Dictatorship that is also strategy-proof.
>
> *Proof.* See appendix.

Note that this theorem does not simply say that constructing deliberative democratic institutions in the proposed way is difficult, but that *it is logically impossible*. This is sufficient to dismiss the current proposal. Because no institutions satisfying minimal requirements of democratic equality are strategy-proof, Bertha can *never* be sure that Althea reveals her true justice-rankings of the values implicit in public reason when she engages in democratic discourse. When inconclusiveness obtains, Bertha can thus never tell if Althea uses public reasons to signal her loyalty to the political

conception of justice, or whether Althea uses public reasons to argue for her private interests. We are, then, back to where we started: the inconclusiveness of public reason causes public reason to fail as an assurance mechanism.

5. Rawls's Solution – Inclusive Public Reason

Rawls recognized that something like the assurance breakdown illustrated in this paper might occur. In response, Rawls first illustrates this potential assurance breakdown by imagining a hypothetical dispute, and then offers his solution to the problem by refining the idea of public reason:[7]

> Suppose that the dispute concerns the principle of fair equality of opportunity as it applies to education for all. Diverse religious groups oppose one another, one group favoring government support for public education alone, another group favoring government support for church schools as well. The first group views the latter policy as incompatible with the so-called separation of church and state, whereas the second denies this. *In this situation those of different faiths may come to doubt the sincerity of one another's allegiance to fundamental political values*.
>
> One way this doubt might be put to rest is for the leaders of the opposing groups to present in the public forum how their comprehensive doctrines do indeed affirm those values. Of course, it is already part of the background culture to examine how various doctrines support, or fail to support, the political conception. But in the present kind of case, should the recognized leaders affirm that fact in the public forum, their doing so may help to show that the overlapping consensus is not a mere modus vivendi (Rawls 1993/2005: 248-249) (emphasis mine).

In this passage Rawls implicitly refers to a distinction between *exclusive public reason* and *inclusive public reason*. Exclusive public reason is what we have been discussing thus far in the paper under the heading "public reason" more generally: on the exclusive view, citizens may only permissibly appeal to a set of reasons shared by all, taken from the political conception of justice. The inclusive view,

[7] This response is also endorsed by Weithman (2010: 329); Weithman (2015: 82).

though, allows "citizens, in certain situations, to present what they regard as the basis of political values rooted in their comprehensive doctrine, provided they do this in ways that strengthen the ideal of public reason itself" (Rawls 1993/2005: 247).

It is important to get clear on exactly what the inclusive view permits. The inclusive view does *not* allow citizens to argue for specific policy conclusions with non-public reasons. Rather, the inclusive view says that when there are assurance breakdowns citizens may appeal to their comprehensive doctrine to *show how it supports the political conception of justice*. That is, the inclusive view of public reason does not say that Althea, if she is a Christian, may appeal to Christian-based reasons when arguing for p_1 or p_2; rather, the inclusive view says that Althea may, in the public sphere, convey why her Christianity endorses the governing conception of justice. The idea here is that, when Althea does this, she gives reason for Bertha to believe that she remains faithful to the political conception of justice. This is important for, as we have seen, Bertha will not be able to conclude Althea's fidelity to the political conception of justice from her public reasoning, for (*i*) inconclusiveness obtains widely, (*ii*) whenever inconclusiveness obtains Bertha has reason to be suspicious of Althea's appeal to public reasons, and (*iii*) there is no logically possible way of structuring deliberation consistent with the ideal of democratic equality such that Althea is incentivized to always give her true justice-based reasons. Because of this widespread problem, hopefully Althea's explaining why her Christianity supports the political conception of justice assures Bertha that she indeed remains faithful to it.

Does Rawls's switch to inclusive public reason solve the assurance problem? No, and for the same reasoning that led to the assurance breakdown with public reason's inconclusiveness in the first place. Recall inconclusiveness says that public reason can support many different policy conclusions. Moreover, we saw that inconclusiveness would be prevalent because people can

reasonably assign different weights to the existing set of public reasons. Something analogous holds for citizens' comprehensive doctrines and conceptions of justice. Comprehensive doctrines contain many different values that govern practical deliberation. Plausibly, these different values can be traded-off and assigned weights in many different ways. Because this is true, comprehensive doctrines can endorse many different political conceptions of justice. Indeed, it is likely that most comprehensive doctrines can support any number of political conceptions of justice – including deeply illiberal ones. Because of this, we see something like a meta-inconclusiveness problem obtain for the inclusive view of public reason: citizen Althea's comprehensive doctrine C will contain values v_1, v_2, v_3,…, v_n. Given how one assigns different weights to v_1-v_n, there will be many different conceptions of justice J_1, J_2, J_3,…, J_n that C could plausibly endorse. Assuming that the governing conception of justice is J_G, when Althea, via inclusive public reasoning, argues that her comprehensive doctrine C supports J_G, Bertha will know that C also consistently supports at least one other $J_{J \neq G}$. Bertha thus has little reason to take Althea's claim seriously that C supports J_G. Althea could plausibly endorse any number of $J_{J \neq G}$'s through C (and, what's worse, illiberal $J_{J \neq G}$'s), and could be merely giving reasons from C to support J_G so she may continue supporting p_1 (which favors her private interest) over p_2 (which is favored by justice).

The Rawlsian might have a response here. In defining what comprehensive doctrines are, Rawls says this:

> [A comprehensive doctrine] organizes and characterizes recognized values so that they are compatible with one another and express an intelligible view of the world. Each doctrine will do this in ways that distinguishes it from other doctrines, for example, *by giving certain values a particular primacy and weight*. In singling out which values to count as especially significant and how to balance them when they conflict, a reasonable comprehensive doctrine is also an exercise of practical reason (Rawls 1993/2005: 59) (emphasis mine).

According to this passage, what it means for Althea to have a comprehensive doctrine implies that her comprehensive doctrine gives a certain "primacy and weight" to values. If we take what Rawls says here literally, then there is no inconclusiveness problem for comprehensive doctrines; for comprehensive doctrines *by definition* come with specific tradeoffs. If Althea's comprehensive doctrine implies one and only one ranking of values, then she will likely be committed to one and only one (let us grant, at least) political conception of justice. As such, Bertha will be able to verify that Althea's inclusive public reasoning is genuine – Althea is not giving a false argument from her comprehensive doctrine by endorsing different tradeoffs than she otherwise would accept, for by hypothesis Althea's comprehensive doctrine entails one and only one ranking of values. If and when Althea says that her comprehensive doctrine supports the political conception of justice, then Bertha will be assured.

In response, we note two different ways of interpreting the term "comprehensive doctrine." With the first interpretation, we take seriously the quoted passage above: a comprehensive doctrine by definition not only includes a bundle of values, but also one unique schema for trading these values off against one another. Call this the *narrow interpretation* of what it means to be a comprehensive doctrine. Rawls also frequently uses "comprehensive doctrine" in a very different way, though: throughout *Political Liberalism*, he frequently refers to Christianity as a comprehensive doctrine. But note, it is clearly false that Christianity broadly construed comes with only one way of weighing the values articulated in Christian doctrine – witness here the difference between the many strands and sects of Catholicism, Protestantism, Calvinism, etc. On this use of "comprehensive doctrine," comprehensive doctrines do *not* come with one assignment of weights and values – instead, they are collections of values, and can be weighed and traded-off according to different standards. Call this the *wide interpretation* of what it means to be a comprehensive doctrine.

Suppose Althea indeed adopts a narrow interpretation of a comprehensive doctrine. But this is not yet enough to allow inclusive public reason to serve as an assurance mechanism. For inclusive public reason to succeed here, Bertha must know Althea's narrow comprehensive doctrine – she must not only know that Althea is a Christian, but she must know the specifics of Althea's particular brand of Christianity: which values from the Bible are included, which are ignored, how they are traded-off against one another, etc. If Bertha does not know this, then we suppose she only knows Althea's wide comprehensive doctrine: that she is a Christian, broadly committed to values v_1-v_n. If Bertha knows Althea's narrow comprehensive doctrine then inclusive public reason serves as an assurance mechanism; if Bertha only knows Althea's wide comprehensive doctrine then inclusive public reason fails as an assurance mechanism.

While it is possible Bertha knows Althea's narrow comprehensive doctrine (and, moreover, the narrow comprehensive doctrines of every other citizen), it is incredibly unlikely that this is so. While most citizens are familiar with the broad tenets of Christianity, Judaism, Mormonism, and the like, very few could cite details concerning all the specific sects of Catholicism and Protestantism, the difference between orthodox and non-orthodox versions of Judaism, and what distinguishes the Apostolic United Brethren from the Fundamentalist Church of Jesus Christ of Latter-Day Saints. Moreover, not only is it unlikely that citizens know such facts, it is implausibly demanding, as a normative requirement, to require they learn them. Because citizens can only be expected to know wide comprehensive doctrines and not narrow comprehensive doctrines, Bertha will not be able to verify if the reasons Althea gives genuinely come from her narrow comprehensive doctrine, or are rather devised to give false allegiance to the political conception of justice. Because of this, inclusive public reason, along with exclusive public reason, fails as an assurance mechanism.

Appendix

The purpose of this appendix is to prove the following theorem:

THEOREM: There is no social welfare function satisfying Pareto, Universal Domain, and Non-Dictatorship that is also strategy-proof.

The theorem is an extension of the Gibbard-Satterthwaite theorem, originally proved independently by Allan Gibbard (1973) and Mark Satterthwaite (1975). The Gibbard-Satterthwaite theorem shows that all resolute voting rules satisfying certain conditions are manipulable. The Duggan-Schwartz theorem extends the Gibbard-Satterthwaite theorem to non-resolute voting rules (Duggan and Schwartz 2000), and the Barbéra-Kelly theorem extends the Gibbard-Satterthwaite theorem to social choice functions (Barbéra 1977; Kelly 1977). To fit the model developed in §4.2, we need to extend the Gibbard-Satterthwaite theorem to social welfare functions specifically.

Why are we extending the Gibbard-Satterthwaite theorem to social welfare functions instead of just using prior versions of the proof? The main thing we are concerned about in this paper is the relationship between different ways of trading off values and the different policy conclusions these different weightings imply. We thus start with different individual *value rankings*, and are then interested in what this means for disagreement over *policy conclusions*. Importantly, *all* different kinds of social choice aggregation mechanisms assume that what is individually ordered and what is socially ordered or chosen are the same unit of analysis: individuals order x, y, and z, and then either x, y, or z is chosen, or x, y, and z are socially ordered in some unique way. If the basic model developed in §4.2 were to simply use existing versions of Gibbard-Satterthwaite that apply to social choice functions or voting rules, then we would move from individual orderings over values to single values being chosen; but here, it's not clear what the relationship between one single value and a policy conclusion is. Or, we would begin with individual orderings over policies and then move to the social choice of one of these policies; but here, the main point of the paper concerning the use of public reasons to assure others in public discourse and the relationship between this assurance mechanism and inconclusiveness has then been forgotten. If, however, we assume unique social orderings of values imply unique policy conclusions, then we can move from individual orderings of values to one social ordering of values and the unique policy commitment it entails. We can then think of individuals as trying to manipulate the resulting social ordering of values so it better aligns with their preferred policy conclusions.

Let $X = \{w, x, y, ..., z\}$ be the set of all public reasons and let $N = \{1, 2, \ldots, n\}$ be the set of all individuals. Every individual i in N has a ranking P_i on X that is total, transitive, and asymmetric, thus mirroring the strict preference relation (no indifference permitted): $x \succ_i y$ denotes "x is ranked above y by i." A social welfare function (SWF) is a function f that takes an n-tuple of individual orderings $\mathbf{P} = \{P_1, P_2, \ldots, P_n\}$ as its argument and states as its value one single social ordering that is

complete, transitive, and reflexive, thus mirroring the weak preference relation: $x \succcurlyeq y$ denotes "x is socially at least as good as y." If $x \succcurlyeq y$ and $y \succcurlyeq x$, then $x \approx y$, which denotes "society is indifferent between x and y." If $x \succcurlyeq y$ and $\neg(y \succcurlyeq x)$ then $x \succ y$, which denotes "society ranks x above y." The social ranking is thus a weak ordering, whereas individual rankings are strict linear orderings.

Pareto: If every individual ranks some alternative x above some other alternative y, then x is socially ranked above y. That is, $\forall i \in N$ and $\forall x, y \in X$, $x \succ_i y \rightarrow x \succ y$.

Universal Domain: The domain of the SWF includes all logically possible *n*-tuples of individual orderings.

Non-Dictatorship: There is no individual whose ranking completely determines the social ranking. That is, $\neg\exists i \in N$ such that, $\forall x, y \in X$, $x \succ_i y \rightarrow x \succ y$.

Strategy-Proof: A SWF is *manipulable* if there exists an *n*-tuple of rankings $\mathbf{P} = \{P_1, P_2, \ldots, P_n\}$ that give the honest justice-orderings $\forall j \in N$, and another *n*-tuple of rankings $\mathbf{P`} = \{P_1, \ldots, P_{i-1}, Q_i, P_{i+1}, \ldots, P_n\}$ where some *i* states dishonest justice-orderings in ranking $Q_i \neq P_i$ and all other individual orderings are equivalent to those in **P**, such that $\exists x, y \in X$ where $x \succ_i y$ in **P** and either (*i*) $y \succ x$ in $f(\mathbf{P})$ and $x \approx` y$ in $f(\mathbf{P`})$; (*ii*) $y \succ x$ in $f(\mathbf{P})$ and $x \succ` y$ in $f(\mathbf{P`})$; or (*iii*) $x \approx y$ in $f(\mathbf{P})$ and $x \succ` y$ in $f(\mathbf{P`})$. In all three cases x fares better than y in the social ordering given the dishonest profile **P`** when compared to the honest profile **P**, where $x \succ_i y$ in the honest profile for individual *i*. If a SWF is not manipulable then it is *strategy-proof*.

The general approach from here is to show that if a SWF is strategy-proof in the manner defined then it also satisfies the Independence of Irrelevant Alternatives. After doing so, the main theorem of the paper reduces to Arrow's impossibility theorem, confirming once again how the Gibbard-Satterthwaite theorem and Arrow's theorem are essentially identical (Reny 2001; Eliaz 2004; Taylor 2005: 69-72; Patty and Penn 2014: 56-57). This general proof strategy also helps vindicate William Vickrey's speculations concerning the relationship between strategy-proofness and independence: "It can plausibly be conjectured that the converse is also true, that is, that if a function is to be immune to strategy and be defined over a comprehensive range of admissible rankings, it must satisfy the independence criterion, though it is not quite so easy to provide a formal proof for this. Immunity to strategy and independence are thus at least closely similar requirements, if not actually logically equivalent" (Vickrey 1960: 518).

Independence of Irrelevant Alternatives (IIA): A SWF satisfies IIA if, for any two ranking *n*-tuples **P** and **P`**, whenever the individual rankings of any two x and y for all $i \in N$ is unchanged between **P** and **P`**, then the social ordering of x and y is the same for $f(\mathbf{P})$ and $f(\mathbf{P}`)$. That is, if how *all* individuals order x and y in **P** is exactly the same as how these individuals order x and y in **P`**, then the social ordering of x and y given **P** must be the same as the social ordering of x and y given **P`**.

Lemma 1. If a SWF function is strategy-proof then it also satisfies IIA.

Proof. We proceed by proving the contrapositive. Suppose SWF *f* does not satisfy IIA. Because IIA does not hold there exist two profiles **P** and **P`** such that all individuals order two arbitrary elements x and y the same in **P** and **P`** respectively, yet $x \succcurlyeq y$ and $\neg(x \succcurlyeq` y)$. We wish to show from this assumption alone that there exist two profiles satisfying our manipulability properties, implying that *f* fails to be strategy-proof. To begin, consider the following sequence:

$f(P_1, P_2, \ldots, P_n)$ yields $x \succcurlyeq y$

$f(P_1`, P_2, \ldots, P_n)$

…

$f(P_1`, P_2`, \ldots, P_{i-1}`, P_i, \ldots, P_n)$ yields $x \succcurlyeq y$ (1)

$f(P_1`, P_2`, \ldots, P_{i-1}`, P_i`, P_{i+1}, \ldots, P_n)$ yields $\neg(x \succcurlyeq y)$ (2)

…

$f(P_1`, P_2`, \ldots, P_{n-1}`, P_n)$

$f(P_1`, P_1`, \ldots, P_n`)$ yields $\neg(x \succcurlyeq y)$

Here, we begin with **P** and replace individual 1, 2,…, *n*'s ordering in **P** with their ordering in **P`**, one individual at a time. Because by hypothesis $f(\mathbf{P})$ yields $x \succcurlyeq y$ and $f(\mathbf{P}`)$ yields $\neg(x \succcurlyeq y)$ there must be at least one point at which replacing an individual's ordering in **P** with their ordering in **P`** swaps the social ordering of x and y from $x \succcurlyeq y$ to $\neg(x \succcurlyeq y)$. Let the first occurrence of such a swap happen with individual *i* and profiles (1) and (2).

Consider now profiles (1) and (2). We wish to show that these two profiles satisfy the relevant manipulability properties, allowing us to conclude that *f* fails to be strategy-proof. Because individual rankings satisfy total, either $x \succ_i y$ or $y \succ_i x$ in (1). Since by hypothesis all individuals order

x and y the same in **P** and **P`**, and since profile (1) contains *i*'s ordering in **P** and profile (2) contains *i*'s ordering in **P`**, if $x \succ_i y$ in (1) then $x \succ_i y$ in (2) and if $y \succ_i x$ in (1) then $y \succ_i x$ in (2).

CASE 1: $x \succ_i y$ in (1) and $x \succ_i y$ in (2).

Suppose *i*'s honest ranking is in profile (2) and *i*'s dishonest ranking is in profile (1). Here, the only difference between (1) and (2) is *i*'s rankings, *i*'s honest ranking says $x \succ_i y$ in (2), the social ordering produced when taking *i*'s honest ranking in (2) says $\neg(x \succcurlyeq y)$, and the social ordering produced when taking *i*'s dishonest ranking in (1) says $x \succcurlyeq y$. By completeness, since $\neg(x \succcurlyeq y)$ in (2) it follows that $y \succcurlyeq x$ in (2), guaranteeing $y \succ x$ in (2). Now either $y \succcurlyeq x$ in (1) or $\neg(y \succcurlyeq x)$ in (1). Suppose $y \succcurlyeq x$ in (1). Then, *f* is manipulable, for $x \succ_i y$ in the honest ranking in (2), the social ordering for the dishonest ranking (1) says $x \approx y$, and the social ordering for the honest ranking (2) says $y \succ x$. Now suppose $\neg(y \succcurlyeq x)$ in (1). Then, *f* is manipulable, for $x \succ_i y$ in the honest ranking in (2), $x \succ y$ in the dishonest ranking in (1), and $y \succ x$ in the honest ranking in (2).

CASE 2: $y \succ_i x$ in (1) and $y \succ_i x$ in (2).

Suppose *i*'s honest ranking is in profile (1) and *i*'s dishonest ranking is in profile (2). Here, the only difference between (1) and (2) is *i*'s rankings, *i*'s honest ranking says $y \succ_i x$ in (1), the social ordering produced when taking *i*'s honest ranking in (1) says $x \succcurlyeq y$, and the social ordering produced when taking *i*'s dishonest ranking in (2) says $\neg(x \succcurlyeq y)$. By completeness, since $\neg(x \succcurlyeq y)$ in (2) it follows that $y \succcurlyeq x$ in (2), guaranteeing $y \succ x$ in (2). Now either $y \succcurlyeq x$ in (1) or $\neg(y \succcurlyeq x)$ in (1). Suppose $y \succcurlyeq x$ in (1). Then, *f* is manipulable, for $y \succ_i x$ in the honest ranking in (1), the social ordering for the honest ranking (1) says $x \approx y$, and the social ordering for the dishonest ranking (2) says $y \succ x$. Now suppose $\neg(y \succcurlyeq x)$ in (1). Then, *f* is manipulable, for $y \succ_i x$ in the honest ranking in (1), $x \succ y$ in the honest ranking (1), and $y \succ x$ in the dishonest ranking (2). ■

Corollary 1. There is no social welfare function satisfying Pareto, Universal Domain, and Non-Dictatorship that is also strategy-proof.

Proof. Since, by Lemma 1, strategy-proofness implies IIA, the proof reduces to Arrow's impossibility theorem. ■ [8]

[8] Though Arrow's original proof assumed both individual rankings and the social ranking are weak orders, the proof goes through if individual rankings are strict linear orderings (total, transitive, and asymmetric) and the social ordering is a weak order (complete, transitive, and reflexive), which is exactly what we have assumed. See Reny (2001); Le Breton and Weymark (2011: 198).

Works Cited

Aumann, Robert. 2000. "Nash Equilibria Are Not Self-Enforcing." In *Collected Papers, Vol. 1*: 615-620. Cambridge: The MIT Press

Barbéra, Salvador. 1977. "Manipulation of Social Decision Functions." *Journal of Economic Theory* 15: 266-278.

Briggs, Rachel. 2010. "Decision-Theoretic Paradoxes as Voting Paradoxes." *Philosophical Review* 119: 1-30.

Chung, Hun. Forthcoming. "The Instability of John Rawls's 'Stability for the Right Reasons.'" In *Episteme*.

Duggan, John and Thomas Schwartz. 2000. "Strategic Manipulability without Resoluteness or Shared Beliefs." *Social Choice and Welfare* 17: 85-93.

Eliaz, Kfir. 2004. "Social Aggregators." *Social Choice and Welfare* 22: 317-330.

Freeman, Samuel. 2007. *Justice and the Social Contract*. Oxford: Oxford University Press.

Gaus, Gerald. 1996. *Justificatory Liberalism*. Oxford: Oxford University Press.

Gaus, Gerald. 2011. "A Tale of Two Sets: Public Reason in Equilibrium." *Public Affairs Quarterly* 25 305-325.

Gibbard, Allan. 1973. "Manipulation of Voting Schemes: a General Result." *Econometrica* 41: 587-601.

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

Kelly, Jerry S. 1977. "Strategy-Proofness and Social Choice Functions without Single-Valuedness." *Econometrica* 45: 439-446.

Kogelmann, Brian and Stephen G.W. Stich. 2016. "When Public Reason Fails Us: Convergence Discourse as Blood Oath." *American Political Science Review* 110: 717-730.

Le Breton, Michel and John A. Weymark. 2011. "Arrovian Social Choice Theory on Economic Domains." In *Handbook of Social Choice and Welfare, Volume Two*. Edited by Kenneth J. Arrow, Amartya Sen, and Kotaro Suzumura: 191-299. Oxford: Elsevier.

MacAskill, William. 2016. "Normative Uncertainty as a Voting Problem." *Mind* 125: 967-1004.

de Marneffe, Peter. 1994. "Rawls's Idea of Public Reason." *Pacific Philosophical Quarterly* 75: 232-250.

Morreau, Michael. 2015. "Theory Choice and Social Choice: Kuhn Vindicated." *Mind* 124: 239-262.

Okasha, Samir. 2011. "Theory Choice and Social Choice: Kuhn versus Arrow." *Mind* 120: 83-115.

Patty, John W. and Elizabeth Maggie Penn. 2014. *Social Choice and Legitimacy: The Possibilities of Impossibility*. Cambridge: Cambridge University Press.

Patty, John W. and Elizabeth Maggie Penn. 2015. "Aggregation, Evaluation, and Social Choice Theory." *The Good Society* 24: 49-72.

Penn, Elizabeth Maggie, John W. Patty, and Sean Gailmard. 2011. "Manipulability and Single-Peakedness: A General Result." *American Journal of Political Science* 55: 436-449.

Quinn, Philip L. 1997. "Political Liberalism and Their Exclusion of the Religious." In *Religion and Contemporary Liberalism*, edited by Paul J. Weithman: 138-161. Notre Dame: Notre Dame University Press.

Quong, Jonathan. 2004. "The Scope of Public Reason." *Political Studies* 52: 233-250.

Quong, Jonathan. 2011. *Liberalism Without Perfection*. Oxford: Oxford University Press.

Quong, Jonathan. 2014: "On the Idea of Public Reason." In *A Companion to Rawls*, edited by David Reidy and Jon Mandle: 265-280. Oxford: Wiley-Blackwell.

Rawls, John. 1963/1999. "The Sense of Justice." In *Collected Papers*: 96-116. Cambridge: Harvard University Press.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

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

Reny, Philip J. 2001. "Arrow's Theorem and the Gibbard-Satterthwaite Theorem: a Unified Approach." *Economics Letters* 70: 99-105.

Satterthwaite, Mark A. 1975. "Strategy-Proofness and Arrow's Conditions." *Journal of Economic Theory* 10: 187-217.

Schwartzman, Micah. 2004. "The Completeness of Public Reason." *Politics, Philosophy, & Economics* 3: 191-220.

Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Stegenga, Jacob. 2015. "Theory Choice and Social Choice: Okasha versus Sen." *Mind* 124: 263-277.

Sunstein, Cass R. and Reid Hastie. 2015. *Wise: Getting Beyond Groupthink to Make Groups Smarter*. Cambridge: Harvard Business Review Press.

Taylor, Alan D. 2005. *Social Choice and the Mathematics of Manipulation*. Cambridge: Cambridge University Press.

Thrasher, John and Kevin Vallier. 2015. "The Fragility of Consensus: Public Reason, Diversity, and Stability." *European Journal of Philosophy* 23: 933-954.

Vallier, Kevin. 2015. "Public Justification Versus Public Deliberation: the Case for Divorce." *Canadian Journal of Philosophy* 45: 139-158.

Vickrey, William. 1960. "Utility, Strategy, and Social Decision Rules." *The Quarterly Journal of Economics* 74: 507-535.

Weithman, Paul. 2010. *Why Political Liberalism?* Oxford: Oxford University Press.

Weithman, Paul. 2015. "Inclusivism, Stability, and Assurance." In *Rawls and Religion*, edited by Tom Bailey and Valentina Gentile: 75-98. New York: Columbia University Press.

Williams, Andrew. 2000. "The Alleged Incompleteness of Public Reason." *Res Publica* 6: 199-211.