Synthese (2019) 196:429–450
https://doi.org/10.1007/s11229-017-1488-7

Constructivism, representation, and stability: path-dependence in public reason theories of justice

John Thrasher[1]

Received: 22 December 2016 / Accepted: 29 June 2017 / Published online: 13 July 2017

Abstract Public reason theories are characterized by three conditions: *constructivism, representation,* and *stability*. Constructivism holds that justification does not rely on any antecedent moral or political values outside of the procedure of agreement. Representation holds that the reasons for the choice in the model must be rationally explicable to real agents outside the model. Stability holds that the principles chosen in the procedure should be stable upon reflection, especially in the face of diversity in a pluralistic society. Choice procedures that involve at least two-stages with different information, as Rawls's theory does, will be path-dependent and not meet the condition of representation since it will not be globally coherent. Attempts to solve this problem without eliminating the segmentation of choice in the procedure will run afoul of constructivism or stability. This problem is instructive because it highlights how public reason theories must evolve in the face of increased concerns about diversity.

Keywords Public reason · Stability · Path-dependence · Social contract · Constructivism

In *A Theory of Justice*, John Rawls argued that the original position is a mechanism by which "the question of justification is settled by working out a problem of deliberation" (1999a, p. 16). Others followed Rawls in adopting this method of justification, classifying their approaches variously as contractarian, contractualists, constructivist, or generally as public reason theories of justice (Rawls 1996; D'Agostino 1996; Gaus

✉ John Thrasher
John.Thrasher@Monash.edu

[1] Philosophy Department, School of Philosophy, History, and International Studies, Monash University, 6th Floor, Menzies Building, 20 Chancellors Walk, Clayton, VIC 3800, Australia

2011). What unifies these approaches is that principles or rules of justice are justified just in case suitably constructed agents in a suitably constructed deliberative setting would choose them—justification depends on reasons for choosing this rather than that set of principles or rules. Public reason theories argue that rational choice in a suitably constructed deliberative procedure can show us the reasons we have for embracing certain principles or rules of justice. This model of idealized choice is normatively important not because we can hypothetically bind ourselves to principles of justice, but because it explicates our reasons for endorsing and complying with particular principles or rules.

Since the publication of *A Theory of Justice* many have objected to this method of justification. The main thrust of these arguments is that hypothetical choice by idealized agents is neither necessary, nor sufficient for justification. Ronald Dworkin argued, "hypothetical contracts do not supply an independent argument for the fairness of enforcing their terms. A hypothetical contract is not simply a pale form of an actual contract; it is no contract at all" (1976, pp. 17–18). In a similar vein, Robert Nozick memorably quipped that contractual justification "isn't worth the paper that it is not written on" (1974, p. 287). More recently, Phillip Pettit has argued that "contract-centered" theories of justification cannot act as the primary justification of principles or rules of justice or morality (1996, 2006). The list goes on.

Despite the scrutiny that contractual and then public reason theories have received, this approach is still dominant in political philosophy. More recently, the criticisms have shifted to concerns with the nature of the idealized choice itself. A larger literature on the benefits and potential dangers of idealization and "ideal theory" in constructivism about justice casts doubts on the nature of modeling justification as ideal deliberation or agreement at all. The literature on this topic has become quite expansive despite the fact that many of the writers on "ideal theory" do not seem to agree about contours of the topic area or about the most important issues at stake (e.g., O'Neill 1987; Mills 2005; Sen 2006; Stemplowska 2008; Valentini 2009, 2012; Simmons 2010; Schmidtz 2011; Estlund 2011; Gilabert 2012; Wiens 2012; Miller 2012; Enoch 2013; Waldron 2013). The problem here is how to relate the principles justified in some procedure to the real world and to what extent the world and people as they are should influence our vision of society as it should be.

More recently a number of theorists have questioned whether public reason theories can remain stable in the face of disagreement in diverse societies. This problem challenged Rawls to develop his later conception of political liberalism and public reason in the first place and it continues to challenge public reason theorists today (Weithman 2010; Gaus 2013). The problem is how commitment to to publicly justified principles can remain stable in a diverse and changing society. One response to this problem is to move in the direction of restricting diversity to a consensus of liberal views, what Quong (2010) calls the "internal" conception of public reason. Others have argued that this consensus approach is inherently fragile and cannot provide the assurance necessary to other citizens to generate stability (Thrasher and Vallier 2015). Gerald Gaus (2011, 2016) goes in the other direction, arguing that public reason theories should seek to be maximally inclusive to diverse viewpoints and values. Regardless, the question of stability in public reason remains a serious one and a problem that will re-emerge below.

Criticism of public reason theories, as we have seen, are not in short supply. Despite this, a fundamental problem has nevertheless been overlooked. This is a problem within ideal theory itself that threatens to push public reason theories into incoherence—failing even on their own terms.[1] Public reason theories of justice require that the selection of principles occur wholly within the idealized choice or agreement procedure. That some construction or "device of representation" can generate a conception of justice without relying on antecedent or external moral and political values is a central feature of public reason theories (Rawls 1996, p. 27). Call this the *constructivism* condition of public reason theories. Yet, simply showing that *some* idealized agents could choose or agree to some set of principles would hardly show that *we* have any reason to endorse of comply with those principles. There must be a relationship between the reasons of the representative choosers in the constructivist model and the reasons of real people. Call this link between the model agents in the constructivist procedure and the reasons of real people outside the model the *representation* condition. If both the *constructivism* and *representation* properties are in place, there is still the question of whether the principles or rules of justice chosen by the representatives will generate their own support and remain stable over time. Call this the *stability* condition.

These three conditions, *constructivism*, *representation,* and *stability*, are essential elements of any plausible public reason theory of justice. Rawls, in particular, spent considerable time refining his account of the last two of these elements in his post-1971 work. Most theories are concerned with the *constructivist* condition and to a lesser extent the *stability* condition. The crucial *representation* condition is less discussed. The problem posed here is that most public reason theories of justice cannot jointly satisfy these three conditions. The reason why is simple, yet very difficult to fix without changing the way we think about some fundamental issues. Or so I will argue.

In the first section, there is a precise characterization of these three essential elements of public reason theories of justice. This is followed by an identification of the problem that creates the conflict, namely the segmenting of choice in the choice procedure into multiple stages with different information. The example of the Original Position is used to illustrate this problem. The rest of the paper is concerned with showing how segmentation creates problems with the three previously articulated conditions and how it is not easily fixed within the context of public reason theories of justice. In the final two sections, some possible solutions are presented before concluding.

1 Constructivism, representation, and stability

Public reason theories of justice are characterized by at least three essential conditions: *constructivism, representation,* and *stability*. Each of these is important and cannot be rejected without significantly altering what is distinctive about the public reason

[1] Throughout, I use "public reason" theories to also refer to many forms of contractualist, contractarian, and constructivist theories generally. It should be clear from the arguments below what theories these criticisms will apply to.

approach. I will argue that the three cannot be jointly satisfied within the context of traditional public reason theory. Before that argument can be made, however, it is important to be clear about these conditions and about the general model of public reason at issue.

Public reason theories are highly abstracted versions of a kind of social contract argument. The goal is to show that members of some society have reason to endorse and comply with the fundamental social rules and principles of that society. Put simply, public reason is concerned with public justification and "the problem of public justification is that of determining whether or not a given regime is legitimate and therefore worthy of loyalty" (D'Agostino 1996, p. 23). The ultimate goal of public reason theories is to show that some political system can meet the challenge Hamilton (1788) raised in *Federalist* no. 1 of whether "men are really capable or not of establishing good government from reflection and choice, or whether they are forever destined to depend for their political constitutions on accident and force."

To show this, public reason theories use a model of justification that has several general parameters that are set differently in different theories. First there are two sets of individuals (N and N^*). The first set is the model choosers constructed in the "device of representation," such as the original position. The second set is composed of real people who populate the society in question. The core problem in public reason theories, what I will call the *representation* condition, is that there be coherently shared reasoning between N and N^* at least as concerns the principles or rules in question. There is also some deliberative setting (M) where the model agents choose some principles or rules and the set of rules or principles that they endorse (R). Given all of this, we can identify a general model of public reason theories (Fig. 1):

> **General Model of Public Reason:** N chooses R in M and this gives N^* reason to endorse and comply with R in the real world insofar as the reasons N has for choosing R in M can be shared by N^*

That choice in this procedure alone can justify those principles or rules is at the heart of why Rawls (1999a, p. 74) describes his theory as "a matter of pure procedural

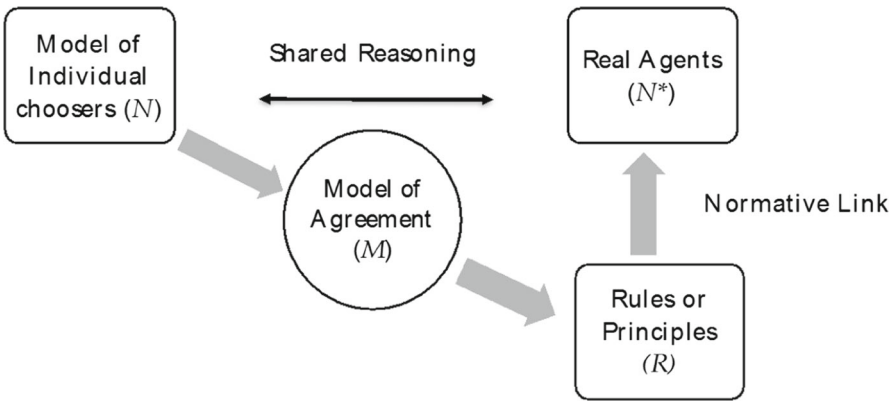

Fig. 1 General model of public reason

justice." This element is also the core of the *constructivism* condition. The procedure tells us what justice is; it does not implement a particular conception of justice. Instead of implementing an antecedently justified conception of justice, the parties in the original position determine what rational individuals would choose in an agreement procedure that models a general point of view (Rawls 1999a, p. 16). For Rawls, the original position does not rely on any external standard of rightness, which the output of the contractual agreement is meant to "mirror." Instead, the original position is constructed "so that the principles that would be chosen, *whatever they turn out to be*, are acceptable from a moral point of view" (1999a, p. 104, Emphasis added). He continues, "[t]hus justice as fairness is able to use the idea of pure procedural justice from the beginning" (1999a, p. 104). This is especially important once the original position becomes the basis for a "free-standing" or "pro-tanto" justification for the principles of a given society in *Political Liberalism* (Rawls 1996, p. 140). Since there will be fundamental disagreement about moral values in a free society, no particular controversial set of values can serve as the basis of our "public moral constitution" (Rawls 1999d, p. 293).

Other contract and public reason theories like those developed by Gauthier (1986), Buchanan (2000), Harman (1975), and Gaus (2011) are similarly committed to *constructivism* insofar as they do not assume any standard of justice prior to agreement. As Rawls puts it, "the conception [of justice] is not regarded as a workable approximation to the moral facts: *there are no such moral facts* to which the principles adopted could approximate" (Rawls 1999c, p. 354, Emphasis Added). Strikingly, he also claims that "apart from the procedure…there are no reasons of justice" (Rawls 1999c, p. 351). Rationally selecting the principles in the original position does not merely help by clarifying the acceptability of the principles chosen, it is essential to their justification.

More precisely, constructivism can be defined negatively as the claim that nothing outside the procedure of agreement or construction can justify or constrain the selection of principles or rules of justice. It can be defined positively as the claim that, once suitably constructed, the agreement procedure is both necessary and sufficient for the justification of some set of principles or rules of justice. Jointly, these two claims comprise the *constructivism* condition:

> **Constructivism** Given a set of suitably constructed deliberators N and a suitably constructed model of agreement M, some set of rules or principles R is justified only if N would agree to R in M.

This condition says two things: (1) that for a given construction of N and M, $R \neq \emptyset$ and (2) the reasons for endorsing R in M are all internal to N and M. Or, put differently, that there is some set of principles or rules that can be justified and justification is self-contained. This characterization leaves open many questions about how to "suitably construct" the agents or the model of agreement. Developing a fully worked account of the deliberators (N), the model of agreement (M), and the target domain of rules or principles (R) sets the parameters for any public reason theory. Precisely how those variables are specified distinguishes one theory from another.

Even this formal and substantively thin definition of constructivism can still illuminate the concept contrastively by showing what cannot count as justification. Any

method of justification that relies on anything aside from the choice or reasons of agents in constructed choice situation will be ruled out. What Christine Korsgaard has called "substantive moral realism," insofar as it relies on the existence of moral facts beyond the parties to an idealized agreement is not constructivist in this sense (1996, p. 35). Many other theories including most forms of consequentialism, intuitionism, and perfectionism will also be ruled out.

The power of this justificatory approach is that it does not rely on any external or controversial moral or political values. Constructivism does not implement any particular conception of justice or morality as such. This makes it especially well suited to justify principles that can "meet the practical requirements of social life" by setting the terms of social cooperation in a society characterized by substantial moral and political diversity (Rawls 1999c, p. 347). This is important because in a diverse society, we can't easily rely on shared moral, political, religious, or even cultural values to anchor our public conception of justice. By using a procedure to generate principles of justice that is independent of any particular sectarian values or commitments those principles will have a justification that anyone in that society will find normatively compelling, regardless of their background.

It is important to note that, in principle, there are no real formal constraints on how the deliberators (N) and the procedure (M) can be modelled in a constructivist theory. Any kind of procedure with any kind of agents can be generated and, insofar as it can be show that there exists some set of principles or rules (R) that can meet the *constructivism* condition, it can be a candidate public reason theory. The next two conditions, however, help constrain the various possibilities implied in the *constructivism* condition. While many different possibilities will meet the *constructivism* condition, substantially less will meet the *representation* and *stability* conditions. In particular, theories that rely on parochial or controversial values and assumptions will be fragile in the face of increased diversity. This is why Rawls was concerned about the importance of pluralism and disagreement in society and why Gauthier, Harsanyi, Buchanan, and Binmore (among others) rely on basic principles and conditions of rationality that they argue everyone can be expected to share to achieve maximal generality and to avoid any fragility in the face of increased diversity.

If constructivism specifies the method of justification, *representation* explicates the relationship between the reasons of the agents in the contractual model and the reasons of real people outside the model. The reasons of the representatives in the choice situation are meant to model *our* reasons; at least the reasons we have that are relevant to ourselves as political agents. The constructivism condition only concerns the reasons of those in the ideal choice situation. Justification, though, is not a "mere proof" (Rawls 1999a, p. 508), nor is it merely reasoning correctly from given or generally accepted premises to conclusions (Rawls 1980, p. 518). Rather, the task is to make explicit the reasoning that connects our standpoint as persons with determinate interests and goals to our standpoint as citizens. This connection between the reasons of agents in the model and reasons of ordinary persons is the link that gives rational choice in the model normative power and what separates it from a mere hypothetical consent.

This connection between the reasons the agents in the contractual procedure have for choosing some set of principles or rules and the reasons we have as real people is

the condition of *representation*. As Rawls (1996, pp. 24–27) argues, the constructivist procedure is a "device of representation" that connects our reasons with the reasons of our representatives in the contractual choice procedure. We can define *representation* more precisely:

> **Representation** Let N be the set of suitably constructed deliberators in a suitably constructed model of agreement M. Let $N*$ be the set of real agents outside the model of agreement M. Let R be a set of possible rules or principles. Let $C_N(\bar{R}) \subseteq R$ be the set of rules chosen by deliberators in N, and $C_{N*}(\hat{R}) \subseteq R$ be the set of rules chosen by deliberators in $N*$. Then, a contractual justification satisfies *representation* if and only if $r \in C_N(\bar{R})$ implies $r \in C_{N*}(\hat{R})$.

Put simply, given the same set of possible rules or principles (R), the choice set of the model agents ($\bar{R}$) implies the choice set of the real agents outside the choice procedure ($\hat{R}$). Representation is the condition that the reasons of the agents in the contractual model must be representable as reasons of agents outside the model and it is a necessary condition for any procedural justification to have normative bearing on real agents. The formal condition of representation will involve, minimally, basic consistency requirements. A choice function will meet this condition if there exists a complete and transitive binary relation ($\succcurlyeq$), such that every admissible choice can be ordered into maximal elements in both choice sets. A, minimal, necessary condition for choice functions to be representable in this sense is that they meet the conditions of expansion and contraction consistency, as we will see below. If it does not, it will be impossible to generate a function that translates $C_N(\bar{R}) \rightarrow C_{N*}(\hat{R})$, which will mean that there is no shared reasoning between the agents in the model and real agents. If representation does not hold, it would be possible to show that agents in the contractual model could choose some option that agents outside the model would have no consistent reason for choosing.

It is important to note here that representation does not mean that either the choice sets or the reasons for choosing need to be exactly the same. If this were a requirement, representation would be reduced to identity.[2] This would defeat the purpose of the constructivist public reason approach. Again, the point of the public reason approach, understood as I have described it, is to make explicit the reasoning that connects our reasoning as distinct individuals with our reasoning as citizens. The point of the deliberative model is not merely, as some have argued, to justify our principles of justice from an impartial or moral point of view. Instead, the goal is to connect two distinct perspectives that we have as persons and as citizens. We cannot be identified with our model reasoners any more than we can be citizens only. But, we need to be able to follow the reasoning of our representative choosers, lest their choice and their reasons become alien to our own. Put another way, the reasoning of our representative choosers needs to be rationalizable to our own.[3] There are many ways to specify this condition, some more strict than others. The weakest is, as I suggested above, to show

[2] I thank an anonymous reviewer for pushing me on this point.

[3] In this context, rationalizability means that the choice set can be understood as a maximal set of a binary ordering of the underlying set of options. This is a basic concept in the theory of rational choice, see: (Bossert et al. 2006).

that a choice function from the model choosers to actual choosers can be generated. This is guaranteed if the original model choice set meets the basic conditions of contraction and expansion consistency. As I will argue below, this condition cannot be met if the choice procedure is segmented and path-dependent.

In addition to constructivism and representation, public reason theories also require a *stability* condition. Stability is essential to Rawls's theory and his way of dealing with stability considerations is what ultimately causes this problem, as we will see in the next section. For now, it suffices to say that stability is the condition that principles or rules chosen according to *constructivism* that meet the condition of *representation* are able to continue to meet those conditions over time. The main idea is that people who have come to endorse certain principles of justice continue to do so in the face of evaluative and doxastic diversity in a given society. More precisely, we can state the stability condition as:

> **Stability** Real agents (N^*) will see themselves as having reasons to endorse and comply with the rules R chosen in the model of agreement (M) in the face of doxastic and evaluative diversity.

Stability is satisfied when members of a society see the principles that govern society as their own and not as impositions. Crucially, principles of justice will only be stable if they are compatible with the deeply held values and conceptions of the good held by those within a society. Too much divergence between conceptions of the good the public conception of the right will make the "strains of commitment" to those public principles unbearable. As such, stability is an essential property of any viable conception of justice meant to govern a diverse and changing society.

These three conditions (constructivism, representation, and stability) are all essential to any public reason theory of justice. The first and second show that principles justified in a particular procedure are normatively compelling to real agents. Stability shows that these principles are robust in the face of evaluative diversity and can serve as the basis of a liberal society. The problem, as we will begin to see in the next section, is that these conditions are often not jointly satisfied. The main problem occurs when the model choice procedure is divided into at least two parts that include different information, as Rawls's original position does. Different information in each stage creates a problem for the *representation* condition and efforts to solve the problem will inevitably undermine either the *existence* or *stability* conditions or both.

2 Choice in stages

The three conditions discussed in the previous section cannot be jointly satisfied in public reason theories with more than one stage of choice, where evaluative or doxastic information differs in each stage. Another way to think of the conditions under which this problem occurs is that certain specifications of the constructivism condition will inevitably conflict with the representation and stability conditions. This is a general problem that applies to any theory that has the features I will describe in the next several sections. Because of its prominence and the detail with which Rawls developed his theory, I will focus particularly on Rawls but this should not be taken to indicate that

this is a problem only for Rawls's theory. Rawls is only used as a particularly vivid example. I will also focus on the procedure developed in *A Theory of Justice*, since Rawls develops that choice procedure in considerable detail. This procedure does not go away in his later work although the stages change somewhat. Nevertheless, as I will show, the problem remains.

Rawls shows that his theory can be stable, but in so doing, segments choice in the original position into two stages, each with different information. This segmentation of choice creates a serious problem. The choosers have different information at each stage and this makes choice in the original position *path-dependent* and, hence, violates the condition of representation. Or so I will argue. In this section we will see how and why Rawls segments choice in the original position and in the next section, I will show why the fact that there is different information in each stage matters.

For Rawls, choice in the original position is made under a thick veil of ignorance. The parties do not know their identities, plans of life, or conceptions of the good. Because of this, they initially have no basis on which to rationally select principles of justice. The solution is to introduce a conception of the good reduced to "bare essential;" what Rawls calls the "thin theory of the good" (1999a, p. 348). This thin theory of the good explicates the rationality of choosing on the basis of primary goods in the original position and allows parties to identify the least well off representative class solely by reference to primary goods (Rawls 1999a, p. 349). In the original position, agents choose the principles of justice based on their thin theory of the good.

This argument is often supposed to be the entire argument for the two principles of justice. Instead, Rawls splits "the argument for the principles of justice into two parts" (1999a, p. 465). The original position argument is only the first part. The second part of the argument introduces a thicker conception of the good as well as the moral psychology of the parties to determine if the principles of justice will be stable, that is, will generate their own support. "Other things equal," Rawls writes, "the persons in the original position will adopt the more stable scheme of principles" (1999a, p. 398). Stability, for Rawls, is *relative* stability. In the original position, the representative choosers compare principles of justice in relation to other competitors in terms of relative stability (Rawls Rawls 1999a, p. 434). It is not enough for a conception of justice to be stable in some sense; it must be *more* stable than other competing options. "Decision in the original position" he writes, "*depends on a comparison*: other things equal, the preferred conception of justice is the *most* stable one (1999a, p. 436).

Rawls argues in Part III of *Theory* that the moral psychology of rational agents lead to congruence between justice as fairness and the good so that individuals develop a sense of justice by seeing justice as bound up with their good (Freeman 2002). Rawls saw this congruence argument as being seriously flawed in his later work and this led him to substantially change his account of justification in *Political Liberalism*. Others have ably documented that change at length and I will not repeat their work here (Weithman 2010; Gaus 2013). The important point to note is that while the original position argument changes somewhat and the account of stability is developed more fully in the later work, the justification of the principles still proceeds in multiple stages in ideal theory. Choice in the original position generates a *pro-tanto* justification for the principles as a freestanding conception of justice. That freestanding conception is then tested under *full* and *public* justification and must be stable under full publicity.

In later work, *inherent stability* becomes *stability for the right reasons*, but the general form of the argument is the same. Principles of justice are rationally chosen in the original position and then those principles are evaluated in terms of relative stability.

Because of this segmentation of the choice procedure, choice in the original position is path-dependent. More generally, public reason theories will not be able to jointly satisfy the conditions of constructivism, representation, and stability in the face of doxastic or informational diversity in the agreement procedure. The principles selected when the stages are ordered one way will almost certainly not be the same principles that are selected if the order is reversed. Segmented choices with different information are prone to inconsistency without some external standard imposed on them (Plott 1973; Sen 1993; Gaertner and Xu 1999; Bandyopadhyay and Sengupta 2006; Poproski 2010; Bossert and Suzumura 2011). Given *constructivism*, however, no external standard of consistency is available to choosers in the procedure. Path-dependence in the choice process makes the choice non-representable and hence undermines the *representation* condition. The cause is informational and evaluative diversity in the two stages of choice. To see why, we need to look at how the presentation of options can affect the rationality of choosing.

3 Diversity and path-dependence

Rational choice can be affected by the presentation of a choice in several different ways. Psychologists and behavioral economists have documented how changing the "frame" or the presentation of options can often change the outcome of choice (Tversky and Kahneman 1981). For instance, people tend to be risk averse when a scenario is presented as choice involving gains and they tend to be risk seeking when a choice is presented as a potential loss, even when the two choices are formally equivalent (Tversky and Kahneman 1986, p. 453). The presentation of options seems to have, in itself, an effect on the valuation of the outcomes. This is what Christian List and Natalie Gold call a "valence framing effect" but we can think of it simply as preferences being affected by the presentation of those options (2004, p. 255). This phenomenon has something to do with the nature of human psychology and while it is important, we can stipulate that the representatives in the original position are somehow able to avoid these kinds of framing effects.

There are, however, other ways that the presentation of options can affect the rationality of a choice. Sometimes the order that options are presented changes the value of those options. This is what Sen (1997) calls "menu-dependence" of choice. This effect is not, strictly speaking, a psychological phenomenon. To see why, consider a case of literal menu-dependence from Luce and Raiffa:

> A gentleman wandering in a strange city at dinner time chances upon a modest restaurant which he enters uncertainly. The waiter informs him that there is no menu, but that evening he may have either broiled salmon at $2.50 or a steak at $4.00. In a first-rate restaurant his choice would have been steak, but considering his unknown surroundings and the different prices he elects the salmon. Soon after, the waiter returns from the kitchen, apologizes profusely, blaming the

> uncommunicative chef for omitting to tell him that fried snails and frog legs are also on the bill of fare at $4.50 each. It so happens that our hero detests them both and would always select salmon in preference to either, yet his response is "Splendid, I'll change my order to steak" (1957, p. 288).

This seems to be an inconsistent choice. The diner's preferences change with the addition of seemingly irrelevant options. After all, the man doesn't like either frog legs or fried snails, so neither option should matter. He seems to have an intransitive preference ordering preferring steak to salmon to steak.

This is not the correct interpretation of this case, however. His preferences are not intransitive. By changing the options available, his preferences thereby changed—his preferences are indexed to the context of a particular menu of options. On a menu that includes salmon and steak, he prefers salmon to steak, but when the menu is enlarged with and option that he dislikes (e.g., frogs' legs), he selects steak. Call the first menu A and the second, enlarged, menu B. $A \subset B$, that is, the first menu is a proper subset of the second menu. The choice in A is salmon because it is preferred to steak. We can express this as Salmon $\succcurlyeq^A$ Steak, which reads "salmon is preferred to steak in menu A." Consistent choice would imply that if the set is expanded to include some irrelevant alternatives, the choice in menu B should also be salmon, but it is not. Instead, the choice in B is steak, an option dominated in the first menu (Steak $\succcurlyeq^B$ Salmon). The diner's preferences have reversed.

Luce and Raiffa explain this apparent inconsistency by arguing that the introduction of the new options is a signal to the man that the restaurant is of a higher quality than he had initially thought. The seemingly irrelevant option provides new information and is, therefore, not irrelevant. The new menu introduces doxastic diversity, changing his evaluation of outcomes on offer and, thereby, his preferences.

Sen uses a more striking example to illustrate the same phenomenon (1993, p. 502). Charles has the option of joining a friend for tea or going home. Given these options and his current beliefs, Charles will choose to have tea. If, after the initial invitation however, Charles's friend tells him that he has changed plans and he now has the option of having tea or cocaine, Charles menu of options has expanded. If we assume that Charles has no interest in cocaine, it looks like we have merely added an irrelevant alternative to the choice menu. The introduction of this irrelevant alternative should not matter to the original decision. Sen argues, however, that the expansion of the menu has added valuable information and is, hence, of epistemic or doxastic value (1993, p. 502). Charles has learned something important about his friend, namely that he sometimes enjoys cocaine with his tea. As Sen puts it, "the expansion of the menu offered by this acquaintance may tell you something about the kind of person he is, and this could affect your decision even to have tea with him" (Sen 1997, p. 753). The lesson to draw is that the choices in these different menus cannot always be treated as equivalent. The introduction of new information has fundamentally altered the choice by changing the information available to the chooser. This shows how slight informational change can cause substantial change in the reasons one has for choosing one option over another.

Doxastic diversity or diachronic doxastic change can have important effects on rational choice. The path through a set of choices can present information to the chooser

in unique ways and can change the attractiveness of various choices. The important point for our purposes is that choices with different information connected to different menus—intra-choice diversity—will tend to be path-dependent. The order of options will matter. Typically, path-dependence has been seen as problem for consistent choice since different options will be choice worthy depending on how those options are presented and there is no guarantee of a globally consistent choice.

This concern is real and it has led many, like Kenneth Arrow, to include Independence of Irrelevant Alternatives (IIA) or some other condition that rules out path dependence at the outset (Plott 1973). But, if we are making decisions with regards to different menus independently, there is no real inconsistency. That is, if our preferences are indexed to particular choice menus (as they were in the example above) the problem seems manageable. All we need is intra-menu consistency, a kind of menu-contextualism for rational choice. To do so, we need to represent preference orderings synchronically based on particular menus. If, however, choices are represented diachronically or choices are made in stages where the past stage determines the options in the next stage, the order of the choices will matter.

Imagine that the diner's choice options were presented differently. Rather than starting with the original condensed menu (A), the waiter starts with B (the expanded menu with snail and frog legs). In this case, the diner would choose steak, not salmon. Now imagine that the waiter comes back from the kitchen in a huff and says, "apologies, it turns out we are out of snails and frog legs." We have gone from $B \rightarrow A$, but the dominant option would still be steak in A. Why? The waiter, in this case, is not giving the diner any relevant new information when he tells him that the kitchen is out of snail or frogs' legs. He already has the information about the quality of the restaurant from the first menu. So, if we start with B and move to $A(B \rightarrow A)$ the choice function is contraction consistent and seemingly path-independent.[4] If, however, we start with A and move to $B(A \rightarrow B)$, the choice moves from salmon to steak and in this direction the choice function does not exhibit expansion consistency and it is clear that the choice is path-dependent.[5] The inconsistency here is understandable, but the choice function is clearly path-dependent, although that fact is hidden if the choice is made in the first direction ($B \rightarrow A$). The ultimate choice the diner will make depends on whether we start with A or B. The path the chooser takes through the decision tree has a substantial effect on whether the man chooses salmon or steak.

In cases like these, path-dependence is not necessarily a result of inconsistency or irrationality of preferences or values. It may, instead, reflect a change of information in the choice circumstances. The structure of the choice situation can, if it introduces doxastic diversity, create path-dependency. Sometimes even the structure of the options alone can insure path-dependence. Consider a classic voting example. Voting systems with certain properties can lead to path-dependent choice when options are compared in multiple stages. In a voting system with three voters {A, B, C} and three issues

[4] What Sen sometimes calls "property α" that $\forall x \in A \subset B \rightarrow [x \in C(B) \rightarrow x \in C(A)]$ (1970, 1993, 1997). I will refer to this property as "contraction consistency."

[5] What Sen sometimes calls "property β," that $[\forall x, y \in C(A) \& A \subset B] \rightarrow [x \in C(B) \rightarrow y \in C(A)]$ (1970, pp. 7–10). I will refer to this property as "expansion consistency."

Table 1 Voting cycle

A	B	C
x	y	z
y	z	x
z	x	y

$\{x, y, z\}$, where each voter ranks their options according to the Table 1, pairwise voting will lead to path-dependent choice.

In this example, the order the options are considered will determine the outcome of the election in majority rule since there is no overall Condorcet winner, that is, an option that can beat any other option it is paired against. If the first vote is over options $\{x, y\}$, x wins because both A and C prefer it to y. If we then vote over options $\{x, z\}$, z wins since B and C prefer z to x. By starting with $\{x, y\}$, voter C is assured to get his best option. Similarly, by starting with $\{y, z\}$, voter A is assured to get her best option and so on. The outcome depends on the order of the options that we consider. Unless there is some external value that the choice is meant to maximize, for instance justice or utility, segmenting choice into stages will often generate the resulting inconsistency and path-dependence (Sen 1993). If choices are only evaluated in terms of their outcomes, that is, without relying on an external criterion of evaluation, path-dependence will almost certainly arise in multi-stage choice procedures (Sen 1997, p. 755).

The main problem with many public reason theories is that that, insofar as they follow Rawls in implicitly or explicitly segmenting choice between two informationally different stages, they will be path-dependent. Path-dependent choice does not meet basic expansion and contraction consistency requirements and, hence, violates *representation*. In the next section we will see why more clearly and assess whether it is possible for Rawls and similar theories to easily solve this problem.

4 Path-dependence and representation

Using Rawls's theory as our example, agents in the original position choose principles on the basis of their thin theory of the good in the first stage of the construction, but in the second stage (with either a thicker conception of the good or under full justification) agents evaluate the principles from the point of view of a determinate conception of the good. In the second stage, the choosers in the procedural construction evaluate the principles from the first stage in terms of whether they will be stable given the costs they impose in terms of their conceptions of the good. Under full justification, some conceptions of justice will ask too much in terms of one's conception of the good and other conceptions of justice will therefore be preferable. The important point is that in the second stage, however we construe it, the agents in the procedure learn something new and important, namely that they have a determinate conception of the good and, to some extent, what it is.

Of course, the reasonable thing to do if one ends up in the second stage with a conception of justice that imposes too high a cost in terms of relative stability is to go back and revise one's selection in the first stage through a process of reflective

equilibrium. The problem with this is that it will conflict with either *constructivism* or *representation* and both are essential. If the choosers can rely on information or considerations that are not available in the first stage, the agreement procedure is either otiose (violating *constructivism)* or path-dependent (violating *representation*). This is the fundamental problem that segmenting choice creates.

To see the problem, consider a toy example of this phenomenon involving the selection of wine. You are tasked with picking the wine for your table. The restaurant has a limited and idiosyncratic wine list with only three options: an excellent, nicely aged Claret, a good California Cabernet, and an acceptable Australian Shiraz. Your ranking of the wines is based solely on the quality and general desirability of the wine. Your initial complete preference ordering is represented in Table 2.

But, this is not the end of the story. Wine isn't free and price matters too. The problem with this ordering is that although it is technically complete, it ignores an important consideration. To rationally select a wine for dinner, you need to know what you will have to give up for it—the opportunity cost of the various options. Spending too much on wine at dinner may mean you have to give up something even more important.

Adding information about cost into the example changes things considerably. The old Claret is being offered tonight at the very reasonable price of $3500 a bottle. The California Cabernet, being significantly less desirable can only fetch a, still very respectable, price of $500 a bottle. Last on the menu is the Australian Shiraz with the attractive price of $40 a bottle. The revised ranking is displayed in Table 3.

Once cost is added, your ranking reverses. This is not a symptom of irrationality but instead an illustration of the prosaic fact that relative cost matters to rational choice. One needs to know what one is giving up before one can truly evaluate different options.

To see why path-dependence creates a problem for rational consistency, consider a diachronic version of the wine example that shares similarity with the selection of principles in the original position. Instead of picking wine in one stage with different menus (one with cost and one without), imagine you select in two stages. In the first, you select the type of wine that you like best. In the second, you select a particular bottle from within that set. The choices in the first stage are Old Claret, California Cabernet, and Aussie Shiraz but the particular bottles and prices are unknown. Let's

Table 2 Wine ordering

Old Claret
California Cabernet
Aussie Shiraz

Table 3 Wine ordering with cost

Ranking without cost	Ranking with cost
Old Claret	Aussie Shiraz
California Cabernet	California Cabernet
Aussie Shiraz	Old Claret

say you prefer Claret to Cabernet to Shiraz. You pick Claret in stage one only to find in stage two that the options this restaurant has are pretty similar in terms of price and quality to the one listed above. They are all very good and very expensive. This is a problem if you can't go back and revise your initial choice.

The point is that adding cost information is not refining or clarifying the choice from the first stage, it is—like in our previous example—changing the choice altogether. Changing menus very nearly reverses one's preferred options when it comes to wine. The situation is even more complicated in the case where one chooses diachronically. Without being able to evaluate the particular options when selecting a type, one cannot, strictly speaking, evaluate their relative costs since it is impossible to know what one is giving up by selecting a particular type of wine. If instead, all of the options are presented at the same time with their relative costs, individuals can make informed decisions. Otherwise, one's most preferred option will potentially lead them to choose something that would be less preferred in the other menu if all the options were known. This happens regardless of which menu you start from.

Although the case presented in this section may seem trivial, the argument extends to choice in constructions like the original position with two stages distinguished by different information. The choice of principles in M_1 (first stage) is a choice that does not take into account the costs of those principles in terms of relative stability in M_2 (second stage). This may seem like a virtue if our task it to merely pick the most attractive principles without considering their costs, but the goal is to pick principles that will be stable as well as attractive. Segmenting the choice of principles in the typical way ($M_1 \rightarrow M_2$) is analogous to choosing types of wine without knowing the relative costs of the actual options available within those types. Introducing cost into the procedure at M_2 the attractiveness of the principles chosen in M_1 will likely change. As in the wine example, this will likely make the choice from ($M_2 \rightarrow M_1$) inconsistent with the choice from ($M_1 \rightarrow M_2$). Because of the informational diversity that segmenting choices with different information and the importance of stability, the selection of principles of justice and their application will be path-dependent, which is merely to say that choice will not be expansion and contraction consistent.

So, multi-stage choice with different information in public reason procedures will be path dependent. There is, however, nothing inherently wrong with path-dependence. It makes inter-menu internal consistency impossible (or accidental), but this is only a problem insofar as inter-menu internal consistency is required for a particular choice. There is no reason, in general, why global consistency (path-independence) should be a necessary standard of rational choice. Internal consistency of choice can be induced on a choice function by imposing an external standard (justice, social welfare, cost) that the overall choice maximizes (Sen 1993). That is, local inconsistency can be made globally consistent if some external standard can impose consistency.

This solution will not work, however, when there is no such external standard to draw on, as is the case in any choice situation characterized by *constructivism*.

Path-dependence without an external standard to induce consistency makes choice in those situations globally inconsistent since they will not meet basic expansion and contraction consistency requirements. Given any subset of the global set of options there is no guarantee that the union of that set with any other set will produce consistent

choices that can be represented as a complete, transitive ordering of outcomes. That is, there is no binary preference ordering that can be used to consistently structure the set of options and, hence, *representation* will fail.

There are many other potentially important conditions that the choice function of agents in the model choice procedure must have for those reasons to represent the reasons of agents outside the model. At a bare minimum, though, the choice within the constructivist model must be representable as a consistent ordering to to represent the choices of real agents outside the model. Without *representation*, there is no reason to think that the reasons of the ideal choosers in the model is normatively linked to our reasons.

The outcome of a path-dependent process is always suspicious. Especially if, like Rawls, one believes the "question of justification is settled by working out a problem of deliberation" it is essential that the deliberative procedure will not likely generate fundamentally misleading results based on arbitrary features like the order alternatives are presented (1999a, p. 16). We recognize that our political institutions are path-dependent and historically contingent in obvious and often unproblematic ways (see e.g., Hardin 1988; D'Agostino 1996; Gaus 2011; Sabl 2012). We would be much more concerned if we saw our fundamental principles of justice as path-dependent in ways that are clearly important.

Moreover, Rawls reiterates the importance of *representation* when he writes that "from the standpoint of the original position, the principles of justice are collectively rational" (1999a, p. 435). They must be explicable to each person's rationality; which is to say that they must be—minimally—representable as a coherent ordering. The reasoning of the parties must be, in the words of Phillip Pettit, "deliberatively accessible" (1996, pp. 294–95). To deny rational explication and *representation* would be to accept that individuals should be subject to the demands of a theory of justice that they, in principle, cannot understand. This is the same as saying the principles have no public rational justification insofar as it is impossible to represent the reasons for their justification (Rawls 1999a, p. 514). The problem can be restated simply as the segmentation of choice in a model deliberative procedure (like the original position) that includes diversity will be path-dependent and this will make *representation* impossible.

The main problem, however, is with stability. Since real agents are not bound to reason in a particular direction, we can be certain that non-representable choice will be unstable in the face of diversity among real agents. Insofar as public reason theories are committed to showing that they can justify some political principles that can act as the basic bedrock of a diverse society, if a theory violates the representation condition, it will not be able to meet the stability condition and will therefore fail as a theory of public reason.

5 Possible solutions

Before turning to the implications of this problem, it is worth considering some possible solutions that the public reason theorist might deploy in order to avoid this problem or lessen its implications. None of these solutions will be entirely attractive and, as

usual with any "impossibility" result, each solution will require that we give up on something else that seems attractive.

The first possibility is to introduce an external standard of consistency, that is, to induce a coherent global structure on the ordering in the model. This may be what Rawls and many commentators have in mind when they invoke the idea of reflective equilibrium, but this escape route is generally blocked by the *constructivism* condition since introducing any external standard will have the effect of weakening *constructivism* and undermining the power and distinctiveness of the public reason approach, especially insofar as the use of any external standard will likely be controversial and make the theory less stable in the face of diversity. Reflective equilibrium, then, can only be useful in how we develop the model, we cannot invoke it within the model itself otherwise we undermine the entire approach and convert it into a form of intuitionism. Or, in any case, this approach will tend to undermine stability.

Along similar lines, we might think it possible to develop some kind of standard of internal consistency that would induce global coherence on choice in the model when choosers in different stages disagree, without introducing any external elements. If so, we might be able to manage disagreement behind the veil without weakening *constructivism* (see, e.g., Muldoon et al. 2014). It is hard to see, however, how this would work without also undermining either the reason for segmenting choice in the first place or introducing an external standard. For instance, imagine that there is one chooser in the original position that we can divide into two based on which stage the chooser is in. Call the choosers in the first and second stage, respectively, C_1 and C_2. Imagine C_2 disagrees with C_1. What is the rational way to resolve this dispute? By hypothesis, C_1 does not have the information available to C_2 and so that information cannot be used to adjudicate the dispute. If both choosers have the same information as C_1 the dispute is rendered non-existent. This makes any choice in the second stage either unnecessary or impossible. Going in the opposite direction has the same effect on the first chooser eliminating the importance of that stage. The simple point is that any principle for adjudicating disputes between the stages will privilege one of the stages and thereby eliminate altogether the need for the other stage. Any solution that did not privilege one of the stages and, thereby make the other stage otiose, would need to use some external value to adjudicate between the stages. But, again, this will either undermine the *constructivism* condition or it will make the initial segmentation unnecessary. In either case, constructivism and segmentation will conflict.

Another, more promising, solution eliminates the segmentation of choice by making all the information available in the first stage. This option involves including all the important details relevant to assessing the principles at the outset. In particular, this would involve including information about the relative compliance costs of different principles. This solves the path-dependence problem and makes the principles more robust with regard to relative stability and the strains of commitment. Knowing the price that we have to pay for principles in terms of our conceptions of the good is essential to rationally endorsing those principles in the same way that knowing the price of wine is essential for making rational selections when we are dining. We recognize that, at some margin, we are willing to limit our pursuit of our conception of the good in order to take advantage of the benefits of collective life under rules of justice that are mutually beneficial.

Insofar as the main reason for segmenting choice is to eliminate information about compliance costs in the selection of principles, including that information in the initial stage of choice will likely make the principles more stable and attractive in one sense, though perhaps this will also limit the aspirational character of those principles or rules. Much of the recent debates about "ideal theory" or "feasibility" is about how much, if at all, certain facts and especially facts about compliance costs should factor into our selection of principles of justice. One important implication of the argument of this paper is that the traditional public reason approaches, exemplified by Rawls, are unstable on this point. Segmentation of choice allows the selection of principles to be done without reliance on compliance facts, but because of this the procedure lacks normative force (because it violates *representation*) or is unstable in the face of information about stability. This instability should lead the public reason theorist to include all the relevant facts in the choice procedure (including facts about compliance) or abandon the approach altogether.

Another concern with introducing all the relevant facts into the initial choice of principles, however, is that doing so might introduce too much informational and evaluative diversity, making any rational choice of principles impossible. This would violate the *constructivism* condition in a different way, showing that *no* construction of a procedure could generate agreement. Rawls recognizes this, noting that a cooperative venture for mutual advantage is "characterized by a conflict as well as an identity of interests" (1999b, p. 130). This diversity of interests may be so great that no agreement is possible. Recent work on introducing diversity into contractual choice suggests strongly, however, that agreement is possible even in the face of substantial diversity of interests and values (Muldoon et al. 2014; Moehler 2014). This problem can be addressed by changing the specification of the procedure so as to ensure at least some agreement. Of course, this can only be done insofar as the new specification is itself consistent with *representation* and *stability*.

There are really two concerns here. One is that the set of rules and principles (R) that can meet the *constructivism* condition is empty. The other is that the set of rules and principles that can meet the *constructivism* condition is non-unique. That is, that there is more than one set of rules or principles can meet whatever test we devise. Rawls was clearly concerned with both problems, but the uniqueness problem explains the changes in the difference principle from "Justice as Fairness" in 1958 to *A Theory of Justice* in 1971. The first version of the second principle did not generate a unique result and this led Rawls to further specify the principle.

In any case, the concern that there is no uniquely rational agreement on a specific set of principles of justice that a diverse set of agents would agree to may be correct. This has led some theorists to use evolutionary models to show that, in a diverse society, individuals can find ways to come to agreement over principles of justice and toleration (Skyrms 1996; Alexander and Skyrms 1999; Bruner 2015). Perhaps surprisingly, Rawls discusses the evolution of the moral psychology necessary for the development of a sense of justice that can induce stability. He writes, "[t]he theory of evolution would suggest that…the capacity for a sense of justice and the moral feelings is an adaptation of mankind to its place in nature" (1999a, p. 440). The attractiveness of this solution will depend crucially on how the evolutionary model works and how *representation* is construed. Evolutionary processes tend to be path-dependent and

will, hence, typically fail the test of representation depending on how they are used in the model.

The most promising version of constructivism that may avoid the problems I have laid out here is Ken Binmore's theory (1998, 2005). He has probably done the most to develop a model that would solve the path-dependence problem with his modification of the original position (1998, pp. 425–435). In his model, two agents who "forget who they are" but who have definite "empathetic" preferences engage in a bargaining game to decide on principles of justice in what Binmore calls the "game of morals" (1998, p. 429). By engaging in this practice, the agents settle on egalitarian principles of justice that would likely meet all conditions above of *constructivism, representation*, and *stability*. In different ways, Gauthier (1986), James Buchanan (1999, 2000), and more recently Moehler (2014) and Gaus (2011) have developed models of contractual agreement that include compliance and stability information in the initial stage. Not all of these approaches would necessarily be consistent with all of the conditions, but insofar as they include much of the relevant information in the first stage, they need to be looked at closer with this problem in mind. Each of these examples show that with a greater specification of the rationality and properties of the contractual agents (N) and greater specificity in the contractual model itself (M), information about compliance need not undermine the possibility of reaching agreement.

Even if the public reason theorist were able to solve the problems discussed here and eliminate path-dependence in the choice procedure, path-dependence may reappear when moving from ideal theory to implementation. The concern with path-dependence developed here is one that arises before any concern with feasibility needs to be addressed, however. In addition, introducing compliance and stability concerns at the outset ensures that the principles selected have already been evaluated in terms of relative feasibility, at least as regards stability. Once the principles are chosen and justified, concerns about path-dependence in their implementation are much less problematic since they do not bear directly on the ultimate justification of those principles.

6 Conclusion

Public reason theories face a serious problem insofar as they segment choice in the deliberative procedure into two stages. When each stage has different information and when the procedure is thoroughly constructivist such that no external standard is used to select the principles or rules in that procedure, choice will be path-dependent. This means that choice will not meet basic conditions of expansion and contraction consistency and, hence, will not be representable to agents outside the procedure. Without the normative link that representation ensures, real agents will have no reason to endorse the output of the procedure as binding on them. Path-dependence thereby undermines the normative force of the entire public reason project.

I have argued that in order to avoid making the choice procedure path-dependent and thereby undermining its normative force, all the relevant information should be included in one stage of choice. Importantly, this means introducing compliance and stability information, but it may mean much more. What this argument shows, if it shows nothing else, is that the relative costs of principles matter when we are choosing

them. But, of course, we knew that all along. The central insight of constructivist and contractual theories of justice is that we all benefit from social life governed by rules of justice, and not only materially—the exact terms of our social life matter. Many possible principles of justice are an improvement on the Hobbesian state of nature but this does not mean that we would want to live under many of these systems. We would hardly see many of them as systems of justice at all. We want to know how much liberty we have to pursue our own good in our own way and how much of this liberty we have to give up to reap the benefits of living in a cooperative venture for mutual advantage. In some systems the benefits are enormous, while in others there may be little benefit at all.

The driving force of the project to develop a public reason theory is to find some way founding our basic political and social principles so that they can meet the bar of public scrutiny. Many prominent public reason theories undermine their own goals by making it necessary to rely on accidental coherence and given the other main property of path-dependent systems, are therefore susceptible to manipulation. Path-dependence, however, is easy enough to avoid in the development of our theories if we know what it involves.

Acknowledgements Thanks to Alexei Procyshyn, Hun Chung, Justin Bruner, Keith Hankins, Leif Wenar, Jerry Gaus, Brian Kogelmann, Chad van Schoelandt, Danny Shahar, and audiences at the University of New South Wales, the University of Canterbury, Seoul National University, The University of Utah, and the University of Arizona for helpful comments on earlier version of this paper and for discussion on the topic.

References

Alexander, J., & Skyrms, B. (1999). Bargaining with neighbors: Is justice contagious? *Journal of Philosophy*, *96*(11), 588–598.

Bandyopadhyay, T., & Sengupta, K. (2006). Rational choice and von Neumann–Morgenstern's stable set: The case of path-dependent procedures. *Social Choice and Welfare*, *27*(3), 611–619. doi:10.1007/s00355-006-0147-6.

Binmore, K. (1998). *Game theory and the social contract, vol. 2: Just playing*. Cambridge: The MIT Press.

Binmore, K. (2005). *Natural justice*. New York: Oxford University Press.

Bossert, W., Sprumont, Y., & Suzumura, K. (2006). Rationalizability of choice functions on general domains without full transitivity. *Social Choice and Welfare*, *27*(3), 435–458.

Bossert, W., & Suzumura, K. (2011). Rationality, external norms, and the epistemic value of menus. *Social Choice and Welfare*, *37*(4), 729–741. doi:10.1007/s00355-011-0568-8.

Bruner, J. P. (2015). Diversity, tolerance, and the social contract. *Politics, Philosophy & Economics*, *14*(4), 429–448.

Buchanan, J. (2000). *The limits of liberty: Between Anarchy and Leviathan. The collected works of James M. Buchanan*. Indianapolis: Liberty Fund.

Buchanan, J., & Tullock, G. (1999). *The calculus of consent: Logical foundations of constitutional democracy. The collected works of James M. Buchanan*. Indianapolis: Liberty Fund.

D'Agostino, F. (1996). *Free public reason: Making it up as we go*. New York: Oxford University Press.

Dworkin, R. (1976). The original position. In D. Norman (Ed.), *Reading Rawls: Critical studies on Rawls' "a theory of justice"*. Palo Alto: Stanford University Press.

Enoch, D. (2013). The disorder of public reason: A critical study of Gerald Gaus's the order of public reason. *Ethics*, *124*(1), 141–176.

Estlund, D. (2011). Human nature and the limits (if any) of political philosophy. *Philosophy & Public Affairs*, *39*(3), 207–237. doi:10.1111/j.1088-4963.2011.01207.x.

Freeman, S. (2002). Congruence and the good of justice. In S. Freeman (Ed.), *The Cambridge companion to Rawls* (pp. 277–315). Cambridge: Cambridge University Press.
Gaertner, W., & Xu, Y. (1999). On rationalizability of choice functions: A characterization of the median. *Social Choice and Welfare*, *16*(4), 629–638. doi:10.1007/s003550050165.
Gaus, G. (2011). *The order of public reason: A theory of freedom and morality in a diverse and bounded world*. Cambridge: Cambridge University Press.
Gaus, G. (2013). The turn to a political liberalism. In M. Jon & D. Reidy (Eds.), *A companion to Rawls* (pp. 233–250). Hoboken: Wiley.
Gaus, G. (2016). *The tyranny of the ideal: Justice in a diverse society*. Princeton: Princeton University Press.
Gauthier, D. (1986). *Morals by agreement*. Oxford: Clarendon Press.
Gilabert, P. (2012). Comparative assessments of justice, political feasibility, and ideal theory. *Ethical Theory and Moral Practice*, *15*(1), 39–56.
Gold, N., & List, C. (2004). Framing as path dependence. *Economics and Philosophy*, *20*(2), 253–277.
Hamilton, A. (1788). Federalist no. 1. In G. W. Carey & J. McClellan (Eds.), *The federalist*, The Gideon Edition, (pp. 1–4). Liberty Fund.
Hardin, R. (1988). *Morality within the limits of reason*. Chicago: University Of Chicago Press.
Harman, G. (1975). Moral relativism defended. *The Philosophical Review*, *84*(1), 3–22.
Korsgaard, C. (1996). *The sources of normativity*. Cambridge: Cambridge University Press.
Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. New York: Wiley.
Miller, D. (2012). *Justice for earthlings: Essays in political philosophy*. Cambridge: Cambridge University Press.
Mills, C. (2005). 'Ideal theory' as ideology. *Hypatia*, *20*(3), 165–184.
Moehler, M. (2014). The scope of instrumental morality. *Philosophical Studies*, *167*(2), 435–451.
Muldoon, R., Lisciandra, C., Colyvan, M., Martini, C., Sillari, G., & Sprenger, J. (2014). Disagreement behind the veil of ignorance. *Philosophical Studies*, *170*(3), 377–394.
Nozick, R. (1974). *Anarchy, state, and utopia*. New York: Basic Books.
O'Neill, O. (1987). Abstraction, idealization and ideology in ethics. *Royal Institute of Philosophy Supplements*, *22*, 55–69. doi:10.1017/S0957042X00003667.
Pettit, P. (1996). *The common mind: An essay on psychology, society, and politics*. New York: Oxford University Press.
Pettit, P. (2006). Can contract theory ground morality? In J. Dreier (Ed.), *Contemporary debates in moral theory* (pp. 77–96). Hoboken: Blackwell.
Plott, C. R. (1973). Path independence, rationality, and social choice. *Econometrica*, *41*(6), 1075–1091.
Poproski, R. (2010). The rationalizability of two-step choices. *Journal of Philosophical Logic*, *39*(6), 713–743. doi:10.1007/s10992-010-9148-0.
Quong, J. (2010). *Liberalism without perfection*. New York: Oxford University Press.
Rawls, J. (1980). Kantian constructivism in moral theory. *The Journal of Philosophy*, *77*(9), 515–572.
Rawls, J. (1996). *Political liberalism. Paperback*. New York: Columbia University Press.
Rawls, J. (1999a). *A theory of justice* (Revised ed.). Cambridge: Belknap Press.
Rawls, J. (1999b). Distributive justice. In S. Freeman (Ed.), *Collected papers* (pp. 130–153). Cambridge, MA: Harvard University Press.
Rawls, J. (1999c). Kantian constructivism in moral theory. In S. Freeman (Ed.), *Collected papers* (pp. 303–358). Cambridge, MA: Harvard University Press.
Rawls, J. (1999d). The independence of moral theory. In S. Freeman (Ed.), *Collected papers* (pp. 286–302). Cambridge, MA: Harvard University Press.
Sabl, A. (2012). *Hume's politics: Coordination and crisis in the "History of England"*. Princeton: Princeton University Press.
Schmidtz, D. (2011). Nonideal theory: What it is and what it needs to be. *Ethics*, *121*(4), 772–796.
Sen, A. (1970). *Collective choice and social welfare*. San Francisco: Holden-Day Inc.
Sen, A. (1993). Internal consistency of choice. *Econometrica*, *61*(3), 495–521.
Sen, A. (1997). Maximization and the act of choice. *Econometrica*, *65*(4), 745–779.
Sen, A. (2006). What do we want from a theory of justice? *The Journal of Philosophy*, *103*(5), 215–238.
Simmons, A. J. (2010). Ideal and nonideal theory. *Philosophy & Public Affairs*, *38*(1), 5–36.
Skyrms, B. (1996). *Evolution of the social contract*. Cambridge: Cambridge University Press.
Stemplowska, Z. (2008). What's ideal about ideal theory? *Social Theory and Practice*, *34*(3), 319–340.

Thrasher, J., & Vallier, K. (2015). The fragility of consensus. *European Journal of Philosophy*, *23*(4), 933–954.
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.
Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *The Journal of Business*, *59*(4), S251–S278.
Valentini, L. (2009). On the apparent paradox of ideal theory. *Journal of Political Philosophy*, *17*(3), 332–355.
Valentini, L. (2012). Ideal vs. non-ideal theory: A conceptual map. *Philosophy Compass*, *7*(9), 654–664.
Waldron, J. (2013). Political political theory: An inaugural lecture. *Journal of Political Philosophy*, *21*(1), 1–23.
Weithman, P. (2010). *Why political liberalism? On John Rawls's political turn*. Oxford: Oxford University Press.
Wiens, D. (2012). Prescribing institutions without ideal theory. *Journal of Political Philosophy*, *20*(1), 45–70.