The Turn to a Political Liberalism

Gerald Gaus

In the preface to *Political Liberalism* (*PL*) Rawls tells us that "to understand the nature and extent of" the differences between *Political Liberalism* and *A Theory of Justice* (*TJ*) "one must see them as arising from trying to resolve a serious problem internal to justice as fairness, namely from the fact that the account of stability in part III of *Theory* is not consistent with the view as a whole" (*PL*, xvii-xviii). Rawls goes on to tell us that the problem of the third part of *Theory* was its claim that a well-ordered society would come to embrace justice as fairness as a "comprehensive philosophical doctrine," but "the fact of reasonable pluralism" shows this to be unrealistic. Recasting justice as fairness to avoid this problem, he tells us, surprisingly "forces many other changes and calls for a family of ideas not needed before" (*PL*, xix). So much is commonplace. There is, however, considerable dispute about almost every aspect of Rawls's "political turn," including whether it was well- or ill-advised (compare Weithman, 2010 and Barry 1995). There is dispute about what Rawls meant by "stability" and whether it is primarily a sociological or a normative-justificatory idea (see Krasnoff, 1998). Some pinpoint chapter IX of *Theory*, "The Good of Justice," as the real source of Rawls's worry (Barry, 1995, 915; Freeman, 2003) while others (Weithman, 2010) stress that close interconnections between all the main elements of part III (chapters VI-IX). And there is extensive dispute about the details of the analysis of part III of *Theory* and in just what way — and what aspects of — *Political Liberalism* sought to overcome them. Rawls apparently thought that almost all the main ideas of political liberalism were the result of fixing *Theory*'s account of stability. While it is manifest that the idea of overlapping consensus is a crucial part of *Political Liberalism*'s response, many also stress the fundamental role of the idea of public

reason (Freeman, 2007) while others emphasize the way in which a concern for legitimacy rather than an account of justice becomes the focus in *Political Liberalism* (Dreben, 2003).

In this chapter I cannot hope to critically survey these debates. Simply to analyze Paul Weithman's (2010) deep and comprehensive recent study of Rawls's political turn would itself require a chapter. Instead, I shall sketch a reading that confirms Rawls's view that the stability argument of part III of *Theory* was crucial for the success of *Theory* as a whole, that it was indeed flawed, and that fundamental ideas of *Political Liberalism* can be traced to the wide-ranging consequences of recognizing this flaw in that argument. The crux of this reading is in line with the main thrust of the fine work of Weithman (2010) and Freeman (2003, 2007), though I do not follow their accounts in all details (which is just as well, since they disagree; Weithman, 2010, 128-9). I then argue, more controversially, that we can find in Rawls's political liberalism at least two (perhaps three) different accounts of the way in which stability considerations enter into justificatory arguments — one repairs the account in *Theory* and is similarly structured, while another pushes political liberalism in a more radical direction. The legacy of the Rawlsian project, I argue, is in developing this latter insight.

1. The Original Position and Stability in *Theory:* The Argumentative Structure

1.1 The first two-stage argument: the justificatory role of stability

The first step in understanding Rawls's political turn is to appreciate that the concern with stability as a basic justificatory consideration was by no means an innovation of political liberalism: it was fundamental to the argument of *Theory*. Rawls was quite explicit that he divided "the argument for the principles of justice into two parts" (*TJ*, 124, 465). The first part is the famous derivation of the two principles of justice via the argument from rational choice in the original position; this derivation requires that part

of the "thin theory of the good" which allows us to identify "primary goods." As is well known, the parties in the original position choose the two principles of justice under a "veil of ignorance" — a range of information that is specific to their own and their society's identity is excluded from the choice situation. Requiring the parties to choose under such conditions helps insure that their choice will be reasonable and not moved by bias (*TJ*, 392). The problem is that without information about what they consider good and their particular plans of life, they do not have a clear basis of choice. Rawls requires that the parties have knowledge of some universal features of good lives, so they know what to aim at (*TJ*, 348-50). The point of part III of *Theory* is to explicate both structural and substantive features of all rational and good plans of life. At this first stage of the derivation all that is required is, as it were, a part of the thin theory: that which specifies certain primary goods — things that rational individuals, "whatever else they want, desire as prerequisites for carrying out their plans of life" (*TJ*, 348). These are liberties, opportunity, wealth, income and the social bases of self-respect (*TJ*, 54). This part of the derivation aims to show that the parties to the original position, exercising their rationality to maximize an index of primary goods, will select the two principles of justice.

Now it is often supposed that this *is* the entire argument from the original position, and once the parties have made their choice their work is done and they can, as it were, fold up shop. Not so (*DP*, 486n). "Persons in the original position," Rawls tells us, must consider whether a well-ordered society founded on justice as fairness will be more stable than alternative conceptions considered in the original position (*TJ*, 398). "Other things equal, persons in the original position will adopt the most stable scheme of principles" (*TJ*, 398). Although the "criterion of stability is not decisive" (*TJ*, 399), if the parties find that a conception is unworkable, this would force a reconsideration of their initial choice (*TJ*, 472, 505).

An ambiguity in Rawls's account of the justificatory role of stability must be noted. The official argument seems to be that of *relative stability*:

> There seems to be no doubt that justice as fairness is a reasonably stable conception of justice. But a decision in the original position depends on a comparison: other things equal, the preferred conception is the most stable one. Ideally we should like to compare justice as fairness with all its rivals in this respect, but as so often I shall only consider the principle of utility (*TJ*, 436).

Somewhat puzzlingly, although Rawls asserts here that there can be "no doubt" that justice as fairness is reasonably stable, and that the hard question is one of relative stability, most of the argument in Part III of *Theory* (the crucial chapter on the "congruence" argument has yet to come) aims to show that justice as fairness is feasibly stable. A few pages after the above quotation he writes:

> These remarks are not intended as justifying reasons for the contract view. The main grounds for the principles of justice have already been presented. At this point we are simply checking whether the conception already adopted is a feasible one and not so unstable that that some other choice might have been better. We are in the second part of the argument in which we ask if the acknowledgment previously made should be reconsidered (§25). I do not contend that justice as fairness is the most stable conception of justice. The understanding to answer this question is far beyond the primitive theory I have sketched. The conception agreed to need only be stable enough (*TJ*, 441).

Near the close of *Theory* Rawls claims in the same paragraph that it has been shown that with respect to stability "the contract doctrine is superior to its rivals" and that the results of part III has been to "justify a conception of justice" by showing that it is "sufficiently stable" (*TJ*, 504-5) — something Rawls had earlier claimed could not be doubted. Most commentators pay little attention to relative stability; I shall follow the

general view of interpreting the main argument for stability in noncomparative terms, understanding it as a "test" that must be passed (Freeman, 2003, 279).

In any event, it must be stressed that in *Theory* the stability test is best understood as what we might call *population stability*. As Weithman (2010, 58, 66) notes, the aim is not to show that the stability test is passed by each and every person — that each and every person will have a stable disposition to act on justice as fairness — but that a well-ordered society has such a general disposition. To be sure, once again Rawls's text is not pellucid; in some passages he rather suggests that the parties, who choose under limited information, would make the same choice when acting on principles of rational choice with full information about their good (*TJ*, 451). The argument for stability can be read as applying to each and every member of a well-ordered society (as does Barry, 1995, 885). However, Rawls acknowledges that even if it is successfully shown that justice as fairness is reasonably stable in a well-ordered society, there may be some citizens "who find that being disposed to act justly is not a good.…in their case just institutions cannot fully answer their nature" (*TJ*, 504). This is important. In this second stage of the argument from the original position parties are not asking whether they (or those whom they represent) will develop the appropriate dispositions, but whether a well-ordered society based on justice as fairness will do so. The parties thus switch from the perspectives of rational self-interested choosers to making a population-level judgment.

1.2 The second two-staged argument: the two elements of stability

Rawls's analysis of the stability of a conception of justice in *Theory* has two parts (*TJ*, 397). First, there is the question whether citizens living under that conception will develop a sense of justice to act on it: whether they will develop desires to act on the principles, and experience the appropriate moral sentiments and natural attitudes regarding them. Rawls thus sketches a moral psychology (*TJ*, §§71-75) that explains how citizens living in a well-ordered society regulated by justice as fairness will develop an

effective desire to act on the principles. Now we might suppose that if we accept this moral psychology the argument for stability has been completed. A society that starts out well-ordered — each accepts justice as fairness and knows that others do and understand its bases — will tend to develop a reinforcing sense of justice in which people come to have an effective desire to act on that conception. What more can be required? As Weithman (2010, 46) shows, Rawls believes that fundamental problems remain. When individuals reason from the "self-interested" view, or the point of view of their own good narrowly defined, they may come to see that acting on their sense of justice is very costly, and so may resent their sense of justice and experience alienation (*TJ*, 295; Weithman, 2010, 53). Thus, considering their good narrowly defined (leaving out the good of acting justly), they may be tempted to injustice. This confronts a well-ordered society with what Rawls called the "hazards of the generalized prisoner's dilemma" — each sees the collective rationality of acting on the principles but is tempted to defect in her own case when recommended by her self-interested point of view (*TJ*, 505, 435, 296; Weithman, 2010, 48). To overcome this hazard, Weithman argues, Rawls sought to show in *Theory* that in a well-ordered society justice as fairness constitutes a Nash equilibrium: "Each member of the W[ell] O[ordered] S[ociety] judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans, *when the plans of others are similarly regulated*" (Weithman, 2010, 64, emphasis in original). Acting justly would then be the best reply to others acting justly.[1]

Rawls's stability argument thus has a second stage: having shown a sense of justice would tend to arise in a well-ordered society regulated by justice as fairness, it must also be shown that the conceptions of the rational good in such a society would be such that people typically are not alienated from their sense of justice. The best life for a typical member of a well-ordered society based on justice as fairness would include a devotion

to, and acting upon, the principles of justice when others do so as well (*TJ*, 382-3). If a society (1) does not encourage a strong sense of justice or (2) encourages conceptions of the good that tempt people away from their sense of justice, the society will fail the stability test as viewed from the original position (*TJ*, 398).

2. Stability in *Theory:* The Substantive Appeal to the Thin Theory

2.1 The thin and full theories related to the question of stability

Part III of *Theory* develops both the "thin" and "full" theories of the good. The thin theory concerns core features of the structure and content of all rational notions of the good life, excluding aspects of the good that appeal to the principles of right or justice. The full theory concerns the good as applied to persons and actions in light of the principles of right (*TJ*, 355, 380ff). As Rawls puts it, the original person can be viewed as a device for developing the thin theory into the full theory by employing the thin theory as the basis for identifying the principles of justice (*TJ*, 382). Once we have knowledge of the principles of right, this constrains and regulates our understanding of what is good (*TJ*, 494-5). In light of the full theory, a life of injustice could not possibly be good, whatever other advantages it may possess. Given this, from the perspective of the full theory, Rawls notes that it is trivially true that maintaining a sense of justice is good for a person, since the full theory is constrained by justice (*TJ*, 498). "Thus what is to be established is that it is rational (as defined by the thin theory of the good) for those in a well-ordered society to affirm their sense of justice as regulative of their plan of life" (*TJ*, 497, 350). Without appealing to the principles of right or justice, and so relying only on general structural and substantive nonmoral features of the good life, Rawls seeks to show that individuals would have strong reasons to affirm their sense of justice under justice as fairness.

2.2 The elements of the thin theory of the good

Before trying to sketch how the thin theory of the good is employed in Rawls's stability analysis, it will be useful to identify its main elements. It goes far beyond the theory of primary goods employed in the first stage of the argument from the original position (*TJ*, 347). The thin theory involves at least six elements.

a. The good as plans of life with a certain structure. The thin theory of the good contains an account of personhood according to which "a person may be regarded as a human life lived according to a plan" (*TJ*, 358). Rawls holds that certain formal principles of rational choice allow a person to identify a "maximal class" of rational plans for her, in which each member is superior to those outside the class but she cannot employ the formal criteria of rational choice to rank the elements within the set (*TJ*, 359, 365). At this point she must employ "deliberative rationality" — a "highly complex [idea], containing many elements" that Rawls does not fully enumerate (*TJ*, 367) — to choose a specific plan. A plan of life consistent with the principles of rational choice and deliberative rationality is a rational plan, and so the person's conception of the good is itself rational (*TJ*, 358-9). A rational interest is one encouraged and provided for by a rational plan (*TJ*, 359); it is from this idea that the account of primary goods is derived (*TJ*, 361). Note that the account of primary goods in *Theory* is thus derived from a conception of the person and the good life.

b. Our social nature. Rawls stresses that "the sociability of humans must not be understood in a trivial fashion" (*TJ*: 458). The account of "goodness as rationality" shows us that there is a maximal class of plans that is rational for us to adopt; we must employ our deliberative rationality to choose one of them. Because "one basic characteristic of humans is that no person can do everything he might do" and so we must choose what abilities to cultivate, the life of each falls short of his full potential

(*TJ*, 458-9). In social life we help complete each other's nature; in a community the members "recognize the good of each as an element in the complete activity the whole of which is intended to give pleasure to all" (*TJ*, 459). Rawls thus affirms the doctrine of our natural "social interest" in the lives of others (Gaus, 1983, chap. 2). It is a consequence of this that the diversity of others' life plans is itself valued in rational plans.

c. Love and Friendship. Agreeing with J.S. Mill, Rawls argues that we possess "natural sentiments of unity and fellow feeling" (*TJ*, 439; Gaus, 1983, 91). As Weithman (2010, 109ff) shows, the thin theory of the good supposes that all members of a well-ordered society seek ties of friendship; we seek relations that can express our attitude of desiring unity with others. This leads members of a well-ordered society with a rational plan of life to participate in associations that promote and express these ties. More deeply, humans experience the natural attitude of love, which is an element of all good lives.

d. Sincerity. It would seem that, as a corollary of our natural desire to live in friendship, we aim at a sort of sincerity in our relations with others (*TJ*, 499-500). Ties of friendship and fellow-feeling render hypocrisy and deception about our actions and motives significant costs (Weithman, 2010, 109).

e. The Aristotelian Principle and its Companion Effect. The above sentiments are at least partly based on the core principle of part III of *Theory*, the Aristotelian Principle, according to which "other things equal, humans enjoy the exercise of their realized capacities (their innate or trained abilities) and this enjoyment increases the more the capacity is realized, or the greater its complexity" (*TJ*, 374). The "exercise of our natural powers," Rawls explains, "is a leading human good" (*TJ*, 374n); this is a

"natural fact" (*TJ*, 376). Rational plans of life thus must take account of this fact which, as Rawls notes, leads to a view of the good that has affinities with the idealist idea of self-realization (*TJ*, 378; Gaus, 1983, 26ff). Failure to develop excellences induces shame (*TJ*, 389). Rawls adds a "companion effect": "As we witness the exercise of well-trained abilities by others, these displays are enjoyed by us and a arouse a desire that we should be able to do the same thing ourselves. We want to be like those persons who can exercise the abilities we find latent in our nature" (*TJ*, 376).

f. The desire to express our nature as free and equal. "Human beings," Rawls tells us, "have a desire to express their nature as free and equal moral persons" (*TJ*, 462). And, he adds, according to the Aristotelian Principle, "this expression of their nature is a fundamental human good" (*TJ*, 390). Again, humans tend to feel shame when they fail to live up to their nature.

2.3 The thin theory and the development of our sense of justice

As noted in section 1.2, the first stage of the argument for stability is that individuals in a well-ordered society regulated by justice as fairness will develop an effective sense of justice — a desire to live up to the principles of justice. Rawls sketches an account of moral development that proceeds through three stages. It is important to realize that this moral psychology is drawn upon by the parties in the original position and affects the choice of principles in the second stage of justification (*TJ*, 405).

In the *Morality of Authority* a child is disposed to act on moral precepts without fear of punishment because of their source in parental authority (or, more generally powerful persons); the precepts are not generally followed because they appeal to the child's inclinations or reason (*TJ*, 408). This first stage presupposes the natural attitude of *love*, for it is the love of, and trust in, the parental authority figures that induces the

disposition to act on their precepts. Although he does not explicitly appeal to it, something along the lines of the companion effect to the Aristotelian Principle holds, since the child wishes to become the sort of person her parents are (*TJ*, 408). The next stage is the *Morality of Authority*, which arises when the child participates in various associations. Here attitudes relating to *fellow feeling* and *friendship* come into play: "once a person's capacity for fellow feeling has been realized in accordance with the first psychological law, then as his associates with evident intention live up to their duties and obligations, he develops friendly feelings toward them...", and this leads to a desire to live up to "the ideals of his station" (*TJ*, 411-2). The companion effect to the Aristotelian Principle is explicitly drawn upon here: witnessing the skills and abilities of others as they do their part, we wish to emulate them (*TJ*, 413). In the *Morality of Principles* ties of friendship still play a role, but rather than wishing to be simply "a good sport" one comes to be devoted to principles that regulate practices beneficial to oneself and those one cares about (*TJ*, 414). The moral sentiments, focusing on principles of justice thus become independent of particular friendships, though Rawls insists that "the sense of justice is continuous with the love mankind" (*TJ*, 417). Perhaps more importantly, acting on Rawls's two principles of justice *expresses our nature as a free and equal rational being*; such expression is an important element of our good (*TJ*, 417). Thus we see that in each stage of the development of the sense of justice critical elements of the thin theory of the good are employed, including the controversial Aristotelian Principle.

2.4 The congruence of the good with justice

The first step in *Theory's* argument for the stability is thus to employ the thin theory to show how an effective sense of justice would arise in a well-ordered society characterized by justice as fairness. But, we have seen, Rawls does not think this is sufficient: if citizens' rational good regularly runs counter to the demands of justice,

people may be tempted to turn their backs on their own sense of justice. To assuage this worry Rawls advances a series of arguments seeking to show how a rational plan of life characterized by the thin theory leads a typical member of the well-ordered society to affirm her sense of justice. As Rawls points out, the "argument is cumulative" and depends on marshaling a variety of considerations; collectively these constitute the overall congruence claim. There is certainly not space here to detail these arguments; Weithman's admirable book painstakingly reconstructs the core arguments, and should be consulted by those wishing to pursue the details. In my view there are four fundamental arguments that comprise the overall congruence claim: (*i*) the argument from the good community, (*ii*) the argument from justice and friendship, (*iii*) the Kantian congruence argument, and (*iv*) the argument from the unity of self.

"It is," says Rawls, "natural to conjecture that that the congruence of the right and the good depends in large part upon whether a well-ordered society achieves the good of community" (*TJ*, 456). The *argument from the good of community* draws on elements *b* and *e* of the thin theory of the good (§2.3). Because we value social life (element *b*), and see the lives of others drawing forth and completing our nature (element *e*), it is part of our good to participate in a society in which others have the freedom and opportunity to flourish (*TJ*, 463). Our participation in a social life with shared ends — a "social union" — is itself a good; a just society itself constitutes a form of social union, "a social union of social unions" (*TJ*, 462). A society regulated by the two principles encourages a diversity of ways of life and, we saw (point *b*) that living in a society characterized by such diversity is part of the rational good. Justice as fairness thus structures a political community in which the excellences of each are brought out by, and compliment, one another "It follows that the collective activity of justice is the preeminent form of human flourishing" (*TJ*, 463), and thus the pursuit of justice constitutes a shared end of the community. The *argument from justice and friendship* also focuses on our social nature. Recall (point *c*) that we have natural fellow feeling; each is united by ties of friendship

with many others in a well-ordered society (*TJ*, 499-500) and so just conduct benefits our friends and loved ones. We have seen that the sense of justice grows out of such love and, indeed constitutes a sort of love of mankind. We want to give justice to those we care about, and we have great difficulty targeting the victims of our injustice (*TJ*, 500); deceiving our friends and fellows in order to gain through injustice by ignoring our sense of justice is especially painful (point *d*).

As Weithman (2010, 182) points out, however, even if successful, these arguments do not show that we desire to act justly for its own sake; we know that in many cases unjust action will set back our good, but we do not know whether just action *as such* is congruent with our good: we do not know whether it is good for us to be "persons who act from the principles of right..." (Weithman, 2010, 190). *The Kantian congruence argument* seeks to overcome this weakness by appealing to a "special feature of our desire to express ourselves as moral persons" identified in point *f* (*TJ*, 503; Weithman, 2010, 190). The desire to express our nature as free moral persons, Rawls argues, simply is (under another description) the desire to act justly (*TJ*, 501; Weithman, 2010, 191). The good of expressing our nature is thus equivalent to a desire to treat our sense of justice as supremely regulative in our life; only a self whose rational plan of life is structured by her sense of justice accommodates this fundamental desire (see also Freeman, 2003, 290ff). Lastly, drawing on his conception of the self as one with a unified plan (element *a*), the argument from a unified self maintains that only a plan of life that accords our sense of justice a regulative role can provide the basis of a unified self. Recall that rational plans have a certain structure and a person is defined as one who lives according to a plan (element *a*). A plan that conforms to the *full* theory of the good (taking the principles of justice as regulative) assures the coherence of the self; it provides an area for our deliberative rationality to exercise itself that accounts for the main elements of the human good. The self is unified not through subservience to a single dominant end such as the pursuit of happiness but through a rationally coherent

plan, fashioned by the deliberative rationality of each in accordance with the principles of right (*TJ*, §85; Freeman, 2003, 295).

3. "The Fact of Reasonable Pluralism"

We have seen that that both stages of *Theory's* stability analysis are based on the thin theory of the good, which is not really all that thin (it might be better characterized as the nonmoral theory of the good, but cf. Barry, 1995, 885ff). Rawls commences the 1996 preface to *Political Liberalism* by proclaiming that its aim is to adjust *Theory's* presentation of justice as fairness to "the fact of reasonable pluralism" (*PL*, xxxvii-xxxviii). What Weithman (2010, chap. 8) has called "the great unraveling" of *Theory's* complex argument for stability has its roots in Rawls's conviction that a diversity of reasonable comprehensive conceptions of the good is "the inevitable long-run result of the powers of human reason at work within the background of enduring free institutions" (*PL*, 4). For *Theory's* stability argument to succeed a free and well-ordered society would have to maintain a consensus (not complete, but overwhelming) on the full theory of the good, which includes justice as fairness as a theory of the right.[2] However, the long-term result of the exercise of reason under free institutions is to induce disagreement on fundamental questions of the good, the nature of the person (Weithman, 2010, 258-9), and our moral natures (whether, for example, the aim of expressing our nature as an autonomous free and equal person is fundamental to our good). The doctrine of the "burdens of judgment" is of decisive importance in the evolution of Rawls's view, for it explains why disagreement about these matters is an enduring feature of a free society. We disagree on these matters because the evidence is often conflicting and difficult to evaluate and even when we agree on the relevant considerations, we often weigh them differently; because our concepts are vague we must rely on interpretations that are often controversial; the manner in which we evaluate evidence and rank considerations

seems to some extent the function of our total life experiences, which of course differ; because different sides of an issue rely on different types of normative considerations, it is often hard to assess their relative merits; in conflicts between values, there often seems to be no uniquely correct answer (*PL*, 56-7). Recognizing the burdens of judgment is constitutive of being reasonable (*PL*, 88-9).

The differences that result are both reasonable and deep. It is not "the fact of pluralism" but the "fact of reasonable pluralism" that motivates Rawls's political turn; pluralism is the result of our best exercise of free practical wisdom (*PL*, 36-7). In stark contrast to differences in rational plans of life in *Theory*, the fact of reasonable pluralism does not suppose the Aristotelian's Principle's implication that our differences are ultimately complimentary, or that we appreciate each other's comprehensive doctrines (Weithman, 2010, 262, 265; but cf. *PL*, 323). We are faced with "intractable struggles" and "irreconcilable conflict" (*PL*, 4, xxviii) of "absolute depth" (*PL*, xxviii).

4. Shallow Political Liberalism: Reasonable Pluralism of the Good

Political Liberalism is a difficult book to explicate. Although much of what Brian Barry claims in his extended review essay is dubious, he seems correct that within the pages of the 1993 edition (and especially within the 1996 edition, which contains a new extended preface and an additional essay — more on that anon), we find inconsistent views; the essays on which *Political Liberalism* is based were written over a number of years, and superseded thoughts appear to be retained along with later ideas (Barry, 1995, 891ff; see also Dreben, 2003, 320). I shall sketch two versions of political liberalism, the "Shallow" and "Deep" Versions (and within the Shallow Version itself I distinguish two formulations). Both can be found in the 1993 text, but I believe that the emphasis in the 1996 edition is more clearly on the Deep Version. In any event, to separately explicate the two versions enhances our understanding of the logic of political liberalism

4.1 Overlapping consensus and stability I: continuity with Theory

The (first formulation of the) Shallow Version informs much of the original 1993 text of *Political Liberalism*, as well as the proto-version of political liberalism we find in *Justice as Fairness: a Restatement* (*JF*). In the Shallow Version Rawls carries over the two-staged derivation of the principles that we examined in *Theory* (§1). First parties in the original position derive the principles of justice under the veil of ignorance using the theory of primary goods, and then the parties check for the stability of the principles by determining whether they can be the focus of a reasonable overlapping consensus (*PL*, 78; *JF*, 88, 181). If not, then "justice as fairness…is in difficulty" — we must go back and see whether the principles can be revised (*PL*, 65-6, 141). However, the parties now cannot go through the reasoning supporting an overlapping consensus; unlike in *Theory* where the theory of the good provides a common basis for checking the stability of the principles that the parties can undertake, overlapping consensus involves different reasons based on different comprehensive doctrines. The parties' task is to determine whether in a well-ordered society of diverse reasonable doctrines there is reason to believe that that the principles can either be derived from diverse comprehensive doctrines, be congruent with them, or at least not conflict with — or at a minimum not conflict "too sharply with"— them (*PL*, 11, 40, 140). Thus, in contrast to *Theory*, rather than demonstrating that the principles *will* be stable, Rawls does not show that an overlapping consensus will occur, but that the freestanding argument allows for it (*PL*, xlvii-viii). Although Weithman (2010, chap. 9) makes out a strong case that, like *Theory*, *Political Liberalism* is concerned with showing that justice as fairness can overcome the "hazards of the generalized prisoner's dilemma," it is perhaps more helpful to stress that Rawls's stability concern in his political liberalism is to show how it is possible for citizens with deeply conflicting comprehensive views to be "wholeheartedly" devoted to a liberal political order (*PL*, xl).

As in *Theory*, the parties are seeking to make a population-level judgment; the overlapping consensus on the political conception should include "all the reasonable opposing religious, philosophical, and moral doctrines likely to persist over generations and to gain a sizable body of adherents…" (*PL*: 15); at least a "substantial majority" of the "politically active citizens" must freely endorse the conception of justice from within their own comprehensive frameworks (*PL*, 38). If, as in *Theory*, the parties are concerned with a population-level stability question, it is not required that each and every reasonable comprehensive doctrine participates in the overlapping consensus; stability requires "sufficiently wide" support (*PL*, 39).

4.2 The two sets model

Thus far the general model of the stability argument in *Theory* carries over into the Shallow Version. Recall that in *Theory*, the case for stability consisted of two stages: first, showing how a sense of justice would develop, and then showing how the thin theory of the good endorses the sense of justice. At one point Rawls affirms that *Political Liberalism* has the same the same structure, except of course that the last step is to show an overlapping consensus supports the sense of justice (*PL*, 140-41n). This no doubt leads some to conclude that the only problem with part III of *Theory* was the congruence argument in Chapter IX, not with the analysis of the sense of justice. But we have seen that the analysis of the sense of justice appealed to most of the now-abandoned thin theory, including the Aristotelian Principle and the good of expressing our natures as free and equal, so the story cannot be as simple as it seems. As we shall see presently, the sense of justice undergoes an important transformation in *Political Liberalism*.

The core stability analysis of Rawls's political liberalisms (Shallow and Deep) proceeds by comparing two sets of values. As Rawls puts it, he supposes that citizens' "overall views have two parts: one part can be seen as to be, or coincide with, the political conception of justice; the other part is a (fully or partially) comprehensive

doctrine to which the political conception is in some manner related (*PL*, 38, xxiii; *JF*, 187; Weithman, 2010, 33; Gaus, 2011). The stability argument is that the values of these two parts *taken together* — the values of the political conception of justice conjoined with the large majority of reasonable comprehensive doctrines — endorse conformity to liberal principles and institutions, and so a well-ordered society based on justice as fairness can be stable (*JF*, 187).

4.3 The political set as freestanding

It is essential to realize that in *Theory* the thin theory of the good was not simply employed in the second stage of the argument from the original position (stability checking), but was employed in the derivation of the two principles of justice in the first stage (§1.1). The account of primary goods was part of the thin theory and, particularly the account of a good life and its structure (element *a,* §2.3). If the fact of reasonable pluralism renders the thin theory of the good unsuitable as the grounds of stability, it certainly renders it unavailable as a supposition in the derivation of the political conception itself. A fundamental aspiration of political liberalism is to free the derivation of justice as fairness from any controversial comprehensive conception by showing that it is free-standing (*PL*, 40, 140). Because justice as fairness can no longer be built up from the reasonably disputable thin theory of the good, it must be built up from fundamental ideas that that "are present in the public culture, or at least in the history of its main institutions and the traditions of their interpretations" (*PL*, 78, 8-9). By commencing with a shared public culture Rawls seeks to assemble the fundamental ideas to be employed in justifying the political set of values and ideas (*JF*, Part I), without grounding in comprehensive doctrines. We can thus appreciate how Rawls's move to a political constructivism, which seeks to construct the set of political values and ideas from the widely shared political culture (*PL*, Lecture III), and so provides the basis of a

freestanding set of political values that is autonomous of comprehensive doctrines (*PL*, 98), is a consequence of his recognition of the fact of reasonable pluralism (*PL*, 38).

4.4 Migrations to the political and decreasing the supporting role of the good

Freeman (2007, 195) notes that ideas associated with the thin theory of the good, which Rawls seems to take away with one hand in *Political Liberalism*, he then gives back with another. One cannot understand the argumentative structure of *Political Liberalism* without tracking the migration (and consequent reinterpretation) of values and ideas that in *Theory* were in the "comprehensive good" set of values into the "political" set of values in *Political Liberalism*. The unifying idea behind this migration is the conception of citizens on which the freestanding, political set of reasons is based. Rawls's fundamental claim is that implicit in our democratic culture is a conception of the citizens who conceive of themselves as free and equal in three senses. First, they understand themselves to possess a moral power to have and revise a conception of the good (*PL*, 30). In explicating this moral power Rawls thus reintroduces a version of goodness as rationality and rational plans of life, but now understood as political ideas (*PL*, 176). Thus, for example, the idea that it is a power of citizens to change their plan of life does not imply that within any comprehensive doctrine such freedom is valued; rather the crucial idea is that our concept of a citizen is such that one's political identity does not change as one's plan does (*PL*, 30). The second sense in which citizens see themselves as free and equal is that they view themselves as possessing valid claims against others, and the third is that they take responsibility for their ends and regulate them according to political justice (*PL*, 32-35). This last clearly concerns a sense of justice, part of the moral powers of citizens (*PL*, 19) — understood now as part of the freestanding set of political values. All this allows Rawls to reintroduce the idea of primary goods, now as derived from our conception of citizens (*PL*, 178-90). Rawls also

reintroduces (now as political values) the social union of social unions (*PL*, 320) and the (political) good of community (*PL*, 201-6).

Two implications of this migration of ideas and values from the conception of the good to the freestanding set of political values should be stressed. First, in evaluating stability, although parties to the original position cannot consider how different comprehensive conceptions support the political conception (since there are many such conceptions), many of the matters that previously were part of the nonpolitical good are available to them as elements of the freestanding political conception. So the parties do have quite a lot of values to appeal to when thinking about stability. Secondly, in the two set analysis of stability, a number of weighty values are included in the political set (*PL*, 139). Now because stability is, as it were, the net effect of the political set and the nonpolitical sets of citizens, because weighty values are included in the political set it is not crucial to show anything like the strong congruence claim of *Theory*. This is why Rawls can correctly say what may at first seem so puzzling, viz. as long as the conflict of the political set with the comprehensive conception *is not too sharp* stability can be achieved (*PL*, 40). Because the political set is weighty, it can bear most of the weight of demonstrating stability so long as there is not a radical conflict with comprehensive conceptions. This feature of political liberalism has, I think, been overlooked: it is not just the case that overlapping consensus replaces the congruence argument — overlapping consensus has a more modest role to play in establishing stability than did the arguments of part III of *Theory*.

4.5 Overlapping consensus and stability II: the individualized, version

In the "Reply to Habermas," added to the 1996 edition, Rawls provides an extended analysis of the relation of justification of the principles of justice to stability, which presents an individualized account of overlapping consensus.[3] While elements of this account were in the 1993 text (e.g., *PL*, 143), the "Reply" clearly sets out an

individualized, rather than a population-focused, analysis of overlapping consensus on a shared conception of justice. Here justification occurs in three stages. The first stage is the freestanding argument from the original position, the argument from the political set. This justification, says Rawls, is only a "*pro tanto*" ("as far as it goes") justification, as it is based only on the freestanding political set (*PL*, 386). The next stage is that of "full justification," which is carried out by individual citizens on the basis of their non-political set of values (their comprehensive conceptions). Here they consider the relation between the implications of the political set and their non-political set; at this stage the justification of the principles of justice "may be overridden once *all* values are tallied up" *(PL*, 386, emphasis added). Note that the justificatory role of overlapping consensus is no longer focused on population-level questions, but concerns each and every reasonable citizen. Unless reasonable citizen Alf affirms the principles on the basis of both sets, the principles are not justified to him. Because the full justification of the principles of justice to Alf depends on the implications of his personal non-political set, it thus is impossible to say whether the principles of justice are fully justified simply by appeal to the argument of the original position.

The last stage of justification is "public justification," "a basic idea of political liberalism" (*PL*, 387). Public justification happens when *all* the reasonable members of political society carry out the justification of the *shared* political conception by embedding it their several comprehensive views" (*PL*, 387, emphasis added). Once public justification occurs there is a common knowledge that each citizen, consulting her own deepest normative convictions, endorses the political conception. We might say all citizens appreciate that the deepest normative and religious convictions of all are reconciled to the public conception: it is public knowledge that the political conception is seen by all as fully justified. Out of public justification comes "stability for the right reasons" (*PL*, 388-9); if achieved, both sets of all citizens (the shared political set and the person's non-political) together endorse a shared political conception, the principles of

justice. This is a demanding account of stability of a particular shared conception of justice: justification requires an overlapping consensus of *all* reasonable citizens, something that was not required of the population-focused account carried over from *Theory*. Because, though, the shared political set does so much of the justificatory work, this claim is not as implausible as it may first appear.

5. Deep Political Liberalism: Reasonable Pluralism of the Right

5.1 The double role of reasonable pluralism

Dreben (2003) thought that the 1996 paperback edition should be considered a second edition of *Political Liberalism*. On his reading the distinctive feature of political liberalism is its principle of legitimacy. On my reading the 1993 edition of *Political Liberalism* contains a second account of liberal stability, which becomes more pronounced in the paperback edition. We can understand this as an implication of a fuller recognition of the ideas that generated the entire political liberal project — the fact of reasonable pluralism and the burdens of judgment. In the introduction to the 1993 edition, Rawls's discussion of reasonable pluralism is focused on the diversity of comprehensive conceptions, but this does not seem to radically infect agreement the political conception; "the political conception is shared by everyone while the reasonable doctrines are not…" (*PL*, xxi). However, in the preface to the paperback edition Rawls stresses that reasonable pluralism and the burdens of judgment apply to the political conception as well:

> In addition to conflicting comprehensive doctrines, PL does recognize that in any actual political society a number of differing liberal political conceptions of justice compete with one another in society's political debate…. This leads to another aim of PL: saying how a well-ordered liberal society is to be formulated given not only reasonable pluralism [of comprehensive conceptions] but a family of reasonable liberal conceptions of justice (*PL*, xlviii).

Rawls thus observes: "The burdens [of judgment] have a double role in PL: they are part of the basis for liberty of conscience and freedom of thought founded on the idea of the reasonable….And they lead us to recognize that there are different and incompatible liberal political conceptions" (*PL*, xlix).

5.2 The principle of liberal legitimacy and public reason

In the introduction to 1996 edition, Rawls states the principle of liberal legitimacy in terms of a principle of reciprocity and justification: "our exercise of political power is proper only when we sincerely believe that the reason we offer for our political action may reasonably be accepted by other citizens as a justification of those actions" (*PL*, xlvi). Now because the fact of reasonable pluralism infects the political set — as we have seen there are many values in the political set, and so the burdens of judgment apply to it — appeal to justice as fairness cannot be required to justify political actions (in matters of basic justice and constitutional essentials; *PL*, 219). It is not the definitively reasonable way to organize and weigh the political values. Consequently, as the implications of the fact of reasonable pluralism for the political set become our main concern, the principle of liberal legitimacy takes center stage. We need to justify our actions to others, and this justification must take into account the fact of reasonable pluralism as applied to the political set. The guidelines for this justification are given by idea of public reason (*PL*, 225-6, 243). In justifying the coercive use of political power on matters of basic justice and constitutional essentials, citizens are to appeal only to conceptions of justice involving reasonable weightings of the political set, along with methods of inquiry which themselves are part of the public culture. Rawls is explicit that the content of public reason cannot be restricted to justice as fairness. "Rather, its content — the principles, ideals, and standards that may be appealed to — are those of a family of reasonable political conceptions of justice…" *(PL*, lii-liii).

5.3 Overlapping consensus and stability III: individualized justification of liberal legitimacy

Note that in the above statement of the principle of legitimacy it is depicted as inherently justificatory, and this justification is owed to other citizens as such.[4] Throughout *Political Liberalism* Rawls argues that the principle of legitimacy seeks to address *each* citizen's reasonable framework to show why they should, given both their sets, endorse a class of political conceptions (*PL*, 137, 143, 224). What Rawls considers the "more realistic" account of an individualized overlapping consensus justification focuses on a "class of liberal conceptions" rather than "a specific conception of justice" (*PL*, 164). Even in "The Reply to Habermas," after applying the three-stage account of justification to a shared specific conception of justice, when Rawls considers whether such overlapping consensus is too "unrealistic to hope for," he moves to legitimacy (*PL*, 392-3). In the end, the individualized account of overlapping consensus most powerfully applies to the justification of a fairly wide set of liberal conceptions of justice and guidelines of public reason conjoined with fundamental aspects of democratic governance (*PL*, 421-33).

6. Conclusion

Theory's wide-ranging use of a not all-that-thin theory of the good to show the congruence of justice and the rational good of citizens supposed that a free liberal society would maintain a consensus on the structure and a much of the substance of a good life. The first corrosive effect of Rawls's conviction that free institutions encourage the growth of reasonable pluralism was to undermine this assumption of a shared liberal theory of the good, and this ushered in what I have called the Shallow Version of political liberalism. This much is clear. The puzzle is in seeing just how Rawls reassembled pieces (and what pieces) of *Theory's* account to produce a "political conception," and to evaluate how much of what was perceived as claims about the good that were too controversial to ground a theory of justice are uncontentious as part of the

liberal political good, which now does the lion's share of the work in the stability argument. And there is the question of whether overlapping consensus and the argument for stability should be seen, as it was in *Theory*, as a population-level, or as an individualized, analysis. As Rawls more fully appreciated how the fact of reasonable pluralism infects not simply ideas about the good, but conceptions of political justice too, political liberalism enters a deeper phase. The aim of treating all as free and equal persons to whom justification is owed is faced with the problem of the indeterminacy of the justification of any specific political conception. It is, I think, Rawls's legacy to present this deep problem to us and show how radically we must revise our political theorizing if we take it seriously. It falls to us to more adequately cope with it.

References

Rawls

PL: 1996. *Political Liberalism*, paperback edition. New York: Columbia University Press.

TJ: 1999. *A Theory of Justice*, revised edition. Cambridge, MA: Belknap Press of Harvard University Press.

DP: 1999 [1989] "The Domain of the Political and Overlapping Consensus." In *John Rawls: Collected Papers*, edited by Samuel Freeman (Cambridge, MA: Belknap Press of Harvard University Press: 473-96.

JF: 2001. *Justice as Fairness: A Restatement*, edited by Erin Kelly. Cambridge MA: Belknap Press of Harvard University Press.

Secondary Works

Barry, Brian. 1995. "John Rawls and the Search for Stability." *Ethics*, vol. 105 (July): 874-915.

Dreben, Burton. 2003. "On Rawls on Political Liberalism." In *The Cambridge Companion to Rawls*, edited by Samuel Freeman. Cambridge: Cambridge University Press: 316-346.

Freeman, Samuel. 2003. "Congruence and the Good of Justice." In *The Cambridge Companion to Rawls*, edited by Samuel Freeman. Cambridge: Cambridge University Press: 277-315.

——. 2007. "Political Liberalism and the Possibility of a Just Democratic Constitution." In his *Justice and the Social Contract*. Oxford: Oxford University Press: 175-214.

Gaus, Gerald F. 1983. *The Modern Liberal Theory of Man*. New York: St. Martin's.

—— 2011. "A Tale of Two Sets: Public Reason in Equilibrium." *Public Affairs Quarterly*, vol. 25 (October): 305-25

Krasnoff, Larry. 1998. "Consensus, Stability, and Normativity in Rawls's Political Turn." *The Journal of Philosophy*, vol. 95 (June): 269-92.

Weithman, Paul. 2010. *Why Political Liberalism? On John Rawls's Political Turn*. New York: Oxford University Press.

Notes

[1] This would not assure stability on justice, for it only shows that acting justly is a possible equilibrium. Because of this we confront a sort of assurance game: we need to be assured that others will play the cooperative equilibrium (Weithman, 2010, 49).

[2] Rawls comes to believe that as a comprehensive doctrine, "the full theory is inadequate" (*PL*, 177n).

[3] Weithman (2010, 335-39) argues for the continuity of the account in "Reply to Habermas" with the overall account in the 1993 version.

[4] In earlier statements, Rawls wrote of what citizens could be expected to "endorse" (*PL*, 137).