Res Publica
DOI 10.1007/s11158-011-9155-1

Rawls and Rousseau: *Amour-Propre* and the Strains of Commitment

Robert Jubb

Abstract In this paper I try to illuminate the Rawlsian architectonic through an interpretation of what Rawls' *Lectures on the History of Political Philosophy* say about Rousseau. I argue that Rawls' emphasis there when discussing Rousseau on interpreting *amour-propre* so as to make it compatible with a life in at least some societies draws attention to, and helps explicate, an analogous feature of his own work, the strains of commitment broadly conceived. Both are centrally connected with protecting a sense of self which is vital for one's own agency. This allows us to appreciate better than much of the literature presently does the requirement for Rawls that justice and the good are congruent, that a society of justice does not disfigure citizens' ability to live out lives relatively unmarked by relations of domination. Some comments on G. A. Cohen's critiques of Rawls are made.

Keywords *Amour-propre* · Finality · Publicity · Rawls · Rousseau

Introduction

Perhaps in part because of its undoubted analytic sophistication, contemporary egalitarian political philosophy is not always particularly interested in its prehistory or in its relationship to its precursors. G. A. Cohen's recent *Rescuing Justice and Equality*, for example, contains no references to Kant or Rousseau despite the centrality of those authors to the evolution of the range of understandings of those two concepts available to us (Cohen 2008, p. 424). Rawls was something of an exception to this. The Kantian inheritance of Justice as Fairness was always both

R. Jubb (✉)
School of Public Policy, UCL, Rubin Building, 29-30 Tavistock Square, London WC1H 9QU, UK
e-mail: R.jubb@ucl.ac.uk

Published online: 28 May 2011

obvious and self-conscious, for example. The publication of Rawls' *Lectures on the History of Political Philosophy* offers us further opportunities to come to better understand Rawls' doctrines in light of his assessments of other figures in political philosophy's canon, and perhaps also to see those other figures in a similarly novel light.

The *Lectures* are careful to make it clear that Rawls originally had various quite particular purposes in composing them. After all, these lectures were originally written as a kind of historical preamble to the lectures which eventually became *Justice as Fairness: A Restatement*, and indeed what Rawls has to say about Mill in particular can probably only be made sense in light of that. It is not clear, though, that Rawls' reasons for finding foreshadowings of his own thought in the past were only narrowly pedagogical. The pedagogical impulse is to be placed in the context of a practically-oriented understanding of the point of political philosophy made only more relevant by the democratic communities in which it was to be exercised. Rawls hoped to 'identify the more central features of liberalism… when liberalism is viewed from within the tradition of democratic constitutionalism' and hence 'interpretations are proposed that seem reasonably accurate to the texts we study and fruitful for my limited purposes in presenting them' (Rawls 2007, hereafter *Lectures*, p. xvii and xviii). We should not expect the *Lectures* to be the most compelling interpretations of the texts and authors they discuss. Rather, we should look to them as explorations, through the historical conversation that took place between both its advocates and its critics, of features of Rawls' understanding of a democratic liberalism. They then become a potentially extremely useful resource for understanding Rawls' own thought. We do not just get to see what he thought of these authors, but the presentation of his conclusions illuminates how the tradition's evolution resulted in his own doctrines. It is not merely that he cannot help revealing himself in what he says about others, but that he deliberately seeks to do so.

Here, I want to focus on Rawls' treatment of Rousseau in the *Lectures*. My interest is in understanding Rawls' presentation of Rousseau in the *Lectures*, and drawing out a particular implication for our understanding of Rawls' own doctrine in light of that. This may seem a rather narrow topic. I do not try and engage with alternative contemporary interpretations of Rousseau, such as those offered by Joshua Cohen or Frederick Neuhouser (Cohen 2010; Neuhouser 2008). Nor do I attempt to adjudicate between these interpretations. My interest here is not in Rousseau interpretation. It is instead in using what Rawls taught undergraduates about Rousseau to understand Rawls better, and so to open up new ways of interrogating his work and that of others. Those new questions will offer potentially fruitful parallels between Rawls and Rousseau as well as for and between other theorists, but I merely indicate, rather than explore, them here. I also hope to encourage others to use the *Lectures* as a resource for understanding Rawls' work more broadly on the basis of the kind of interpretative maxims I lay out here. What little has been written on them seems to me to misunderstand them. Michael L. Frazer has argued that the *Lectures* should be understood as an exercise in humility by Rawls and so read as guided by a maxim of charity (see Frazer 2010). This seems wrong. Making that point, though, is subsidiary to the main one about the best way

to understand Rawls and the challenges facing political theorists more generally through the discussion of Rousseau in the *Lectures*.

Roughly, the idea is that if we are to understand the *Lectures* as explorations of the tradition of democratic constitutionalism, then Rawls' insistence on reconciling Rousseau's idea of *amour-propre* with a just political order should be seen as pointing to an important feature of his understanding of democratic constitutionalism. More, that Rawls presents Rousseau as able to achieve this reconciliation suggests that he thinks we ought to understand this feature of democratic constitutionalism in much the way that Rousseau does. Whereas Locke, Hobbes, Hume, Mill, and Marx have problems with their theories pointed out, Rousseau does not. Thus, in this paper I try and do two things. First, I try and identify the relevant feature of Rawls' own theory he highlights by emphasising the significance of *amour-propre* in interpreting Rousseau. Second, I explore what we might learn about that feature of Rawls' thought by comparing it with Rousseau's idea of *amour-propre*.

I argue that by pointing to the significance of Rousseau's idea of *amour-propre* Rawls is emphasising the use he puts formal constraints on the concept of right to when arguing against utilitarianism. I will refer to this feature as the 'strains of commitment' although this is not quite the same as Rawls' own use of the term.[1] My reason for identifying this group of ideas as those Rawls' treatment of Rousseau is supposed to point out is the similarity in Rawls' presentation of them. Both are presented not only as being satisfied by the proper public affirmation of all individuals' self-worth, but as importantly relating to agents' conceptions of themselves as purposive actors. I then argue that the strains of commitment in this broad sense have not always had their significance appreciated.

The implication of Rawls' comparison is that societies that do not conform to his principles of justice will be disfigured by the kinds of pathologies that Rousseau identifies with corrupt *amour-propre*. To what extent that is true, to what extent it is true because of reasonable behaviour, and to what extent it matters are not questions which seem to me to have been much considered, let alone satisfactorily answered. As Ivar Labukt puts it, discussion of Rawls' claim to have shown his theory to be superior to utilitarianism 'has almost exclusively been focused on narrow and somewhat technical questions about what it would be rational to want behind a veil of ignorance' rather than on the 'role played by considerations of feasibility or practicability' (Labukt 2009 pp. 201–202). Labukt is only interested in the first of my three questions though. Whilst we reasonably want to know whether societies that do not conform to Rawls' principles of justice will be disfigured by the kinds of pathologies Rousseau identifies with corrupt *amour-propre*, that does not tell us whether individuals have a claim in justice against being exposed to such pathologies. The claim about feasibility, about whether a society of Rawlsian justice can perpetuate itself, cannot by itself justify justice giving weight to avoiding these

[1] Rawls uses the strains of commitment to refer only to considerations of finality, referring to considerations of publicity under the heading of stability. I prefer to use the broader term since both are to do with the excessive demands a political system might make and because stability has unfortunate connotations of a *modus vivendi*.

pathologies. Perhaps justice should be done though the heavens may fall. We need to answer questions about why individuals are entitled to protection against a society which generates the sorts of pathologies associated with corrupt *amour-propre*. By developing the framework in which such questions can be productively posed, I hope to begin to the task of providing answers to them. In doing so, I offer new ways of understanding disputes around G. A. Cohen's critiques of Rawls.

This paper has three Sections. The first lays out Rawls' interpretation of Rousseau and locates it amongst the available options. The second Section provides an interpretation of Rawls' insistence on the possibility of the society envisaged in *On The Social Contract* whilst retaining the doctrines of *The Discourse on the Origin of Inequality* as a condition of the coherence of Rousseau's political philosophy.[2] It argues that the insistence shows the centrality of Rawls' own way of ensuring that a just society could avoid systematically disfiguring its citizens' lives to his project. The third Section attempts to further motivate Rawls' constraints, and to draw attention to the way that this motivation for them has often been missed. Obviously, the paper is more aimed at Rawls(ian) than Rousseau(vian) scholars, but I hope that it is of interest to both: perhaps in trying to expose connections between the two, we will come to understand better the difficulties of '[f]ind[ing] a form of association which defends and protects with all common forces the person and goods of each associate, and by means of which each one, while uniting with all, nevertheless obeys only himself and remains as free as before' (*SC,* 1.6.4).

Rousseau's Realistic Utopia

Rawls made it clear that the *Lectures* are not best understood as straightforward accounts in the history of ideas. In particular, Rawls seemed to be using his discussion of historical figures in the *Lectures* to make points about the advantages of the structure of his own doctrine. He does this by using problems in their views to suggest the developments he sees his own theory as having made. His criticisms of Hobbes for lacking a theory of the reasonable distinct from a theory of the rational, for example, can be seen in this light, as can his concerns about Locke's use of a pre-existing moral doctrine (*Lectures*, p. 87, p. 104, p. 112). We might expect then that Rawls' discussion of Rousseau in the *Lectures* will have a critique which highlights the appropriateness of a particular feature of Justice as Fairness. However, there seems to be no such critique.

What preoccupies Rawls in his discussion of Rousseau is the problem of reconciling the 'dark and pessimistic' *Second Discourse* with the 'sunnier' *Social Contract* (*Lectures*, p. 193). He sees the truth of the claim in the *Discourse* that human nature is good but has been corrupted as depending on whether political institutions can satisfy both 'the principles of political right' and meet 'the requirements for institutional stability and human happiness' (*Lectures*, p. 206). Thus, the question becomes whether the *Social Contract*'s vision of possible

[2] Citations of *On The Social Contract* (hereafter *SC*) will be by book, chapter, and paragraph, but I will refer to page numbers for *The Discourse on the Origin of Inequality* (hereafter *Discourse*).

legitimate political institutions can be made to work with the *Discourse*. Rawls concludes that it can. Crucial to his argument is the claim that Rousseau understands *amour-propre* as naturally being 'a need which directs us to secure for ourselves equal standing along with others' (*Lectures*, p. 198). This is a need which the society of the *Social Contract* can meet by replacing personal dependence with interdependence on terms all can accept. Under the social contract, the pathologies of *amour-propre* are dealt with and so the terrible state of other developed societies will be left behind.

Rawls acknowledges that his account of *amour-propre* and the possibility of a non-destructive form of it draws heavily on work by Nicholas Dent and Frederick Neuhouser. The central role of *amour-propre* in humankind's descent into vice provided in the *Discourse* is clear. The question is whether it inevitably brings about the world of base equality where 'Subjects have no other Law left than the will of their Master, and the master no other rule than his passions' that the *Discourse* ends with (*Discourse*, p. 185). Dent, for example, has argued that *Emile* shows that 'the central demand of *amour-propre*… is capable of being met providing we are clear about what is our due from others—not servility and fawning adulation but a position among men of common regard and common respect' (Dent 2005, p. 105). Nor is this only possible for isolated individuals, but for whole societies. What satisfies *amour-propre* is a position of *common* regard and *common* respect and so 'the demands of any one individual's *amour-propre* can be met consistently with those of each and every other person' (Dent 2005, p. 105). Rawls' wide reading of *amour-propre* is then, not widely at variance with the broader literature. Although law alone cannot guarantee all a position of common regard and common respect, as Neuhouser notes, '*equality of respect* as citizens' and the 'safeguarding of fundamental interests' can secure a 'recognized standing within the community that is itself a form of respect' (Neuhouser 1993, pp. 390–391).

Having thus positioned Rousseau as not just a 'critic of culture and civilization', of 'the deep-rooted evils of contemporary society' and its members' 'vices and miseries', but one who tries to 'describe the basic framework of a political and social world in which they would not be present' (*Lectures*, p. 192), Rawls then goes on to reconstruct Rousseau's description of such a framework. Obviously, the terms of the social contract are crucial here. Rawls sees the contract as based on four assumptions made by Rousseau. First, those cooperating aim to advance fundamental interests importantly connected 'with the love of self in both of its proper natural forms' (*Lectures*, p. 217). *Amour de soi* generates interests in perfectibility and free will and *amour-propre* an interest in 'having a secure standing… as an equal member of our social group' (*Lectures*, p. 218). Second, although individual relations of dependence are terrible and an important source of the perversion of *amour-propre*, social cooperation is both 'necessary and mutually advantageous' (*Lectures*, p. 218). Third and fourth, we are equally capable of acting to advance the equal interests we have in light of our *amour de soi* and have a similarly equal capacity for and interest in acting justly, respectively. Hence, mirroring Rousseau's own formula, the question becomes 'how, then, without sacrificing our freedom, to unite with others to secure the fulfilment of our

fundamental interests, and to guarantee the conditions for the development and exercise of our capacities' (*Lectures*, pp. 219–220).

Understanding the question then dictates the proper answer: 'the total alienation of each associate with all his rights to the whole community' (*SC,* 1.6.6). Rawls then glosses Rousseau's comments on that single clause, noting that the equality of the alienation and the interests in freedom mean this is not a licence for illiberal interference; reading the claim that the 'union is a perfect as it can be' (*SC,* 1.6.7) as an insistence on the social contract's derivation of principles of ultimate political right; and observing the way in which equal alienation removes relations of personal dependence (*Lectures*, pp. 220–222). The general will is not then 'the will of an entity that in some way transcends the members of society' but 'a form of deliberative reason shared and exercised by each citizen as a member of the corporate body… that comes into being with the social compact' (*Lectures*, p. 224, p. 227).

This form of deliberative reason is based on citizens' common interests, interests based not on 'people as they actually are in a society marked by extremes of inequality… with the resulting evil of domination and subjection' but rather what Rawls describes as 'Rousseau's conception of the person as a normative idea' (*Lectures*, p. 226, p. 228). Since these interests are common, they form a kind of public reason: they make up a conception of how 'citizens who, as a collective body, exercise final political and coercive power over one another' ought to reason (Rawls 1993, p. 214). That publicity, which requires mutuality, delivers justice and a kind of equality since each thinks of every other as they think of themselves.

Barring some final remarks on Rousseau's views on equality and discussions of the Legislator and the infamous 'forced to be free' passage, this is the broad outline of Rawls' reading of Rousseau in the *Lectures*. The rather intolerant demands for tolerance, for example, do not get mentioned and Rawls clearly takes extensive pains to make Rousseau seem like an egalitarian liberal; albeit, an egalitarian liberal of a particular stripe, but one who would make a sort of sense in the contemporary scene. That alone seems to me a service, but that is not all that there is to remark on in Rawls' reading of Rousseau. The problem Rawls sets himself in discussing Rousseau, of whether the *Social Contract* and the *Discourse* can be made compatible, is answered in the affirmative. By giving its citizens a reciprocally-granted equal status, the society of the social contract tames *amour-propre*. The question then becomes, if there is not a problem in democratic liberalism that Rawls has a better solution to than Rousseau does, what is Rawls' discussion in the *Lectures* supposed to teach us about the contours of that tradition?

Amour-Propre and the Strains of Commitment

We can begin to understand the point of Rawls' discussion of Rousseau by looking more closely at his characterisation of *amour-propre* and in particular comparing it with his explanation of why the parties in the original position would choose his two principles over utilitarianism. See the two passages below:

> In its natural… form… *amour-propre* is a need which directs us to secure for ourselves equal standing along with others and a position among our associates in which we are accepted as having needs and aspirations which must be taken into account on the same basis as those of everyone else. This means that on the basis of our needs and wants we can make claims which are endorsed by others as imposing rightful limits on their conduct. Needing and asking for this acceptance from others involves giving the same to them in return (*Lectures*, pp. 198–199).

> What the principle of utility asks is… a sacrifice of [life] prospects. We are to accept the greater advantages of others as a sufficient reason for lower expectations over the whole course of our life. This is surely an extreme demand. In fact, when society is conceived as a system of cooperation designed to advance the good of its members, it seems quite incredible that some citizens should be expected, on the basis of political principles, to accept lower prospects of life for the sake of others (Rawls 1971, hereafter *Theory*, p. 178) .[3]

It seems clear that the discussion of how utilitarianism would fail to meet the strains of commitment is the obverse of the set of demands that the natural form of *amour-propre* makes. If my 'needs and aspirations must be taken into account on the same basis as those of everyone else', then it would be hard 'to accept the greater advantages of others as a sufficient reason for lower expectations over the whole course of my life'. It does not do justice to the 'rightful limits on their conduct' which those needs and aspirations impose. Utilitarianism treats people as equals as bearers and producers of utility, but treating them as equals in that sense is not the sense which Rawls' reading of natural *amour-propre* demands. Doing so does not address the problem of their standing, their status amongst their fellows, or their needs and aspirations.[4] Utilitarianism notoriously may justify slavery, and slaves lack equal status, and neither is respecting the utility I and others get from my projects the same as respecting those projects themselves.

It is not particularly surprising that Rawls thinks that Rousseau ends up ruling out utilitarianism, given Rawls' own hostility to it. Nor is it obviously an odd reading of Rousseau. As Rawls mentions, it looks like Rousseau's understanding of what the general will aims at is a shared, rather than summed, good (*Lectures*, p. 230, *SC*, 2.1.1, 2.3.2). What is distinctive about the two passages above is that Rawls' understanding of what would satisfy natural *amour-propre* seems to parallel his own reasoning for the rejection of utilitarianism. The features which would make a society publicly governed by utilitarian principles unbearable for some are markedly similar to Rawls' understanding of the distinctive features of a society marked by corrupt *amour-propre*. What makes those societies unable to meet the strains of commitment is that they do not respond adequately to the demands of something

[3] That this is strictly a consideration of stability, I think, shows that it is reasonable to group both considerations of finality and publicity under the idea of the strains of commitment.

[4] As Samuel Freeman puts it, '[e]qual consideration in a hypothetical decision process is not a good' in the sense that 'the resources and opportunities that enable people to achieve happiness and lead a good life' are (Freeman 1994, p. 329).

like natural *amour-propre*. Rawls' emphasis in the *Lectures* on the importance of being able to reconcile the *Discourse*'s 'darkly pessimistic' description of the evolution of an obsession with status that ultimately leaves people 'capable of living only in the opinion of others' with the *Social Contract*'s 'sunnier' view then comes down to a way of stressing the importance of the strains of commitment (*Lectures*, p. 200, p. 187, p. 193). By insisting that a society in which *amour-propre* is satisfied is possible and desirable, Rawls is insisting that a society in which the demands of the strains of commitment are satisfied is also possible and desirable.

Bearing in mind Rawls' distinctive purposes in the *Lectures*, this may also explain various other features of Rawls' discussion of Rousseau there. For example, the absence of any discussion of what Rousseau has to say about religion in the *Social Contract* avoids bringing up the way in which Rousseau drew the conditions under which one could stand in relations of equality to one's fellow citizens much more narrowly than Rawls did. After all, as a contemporary liberal Rawls could hardly have been happy with the thought that the separation of Church and State makes 'good polity impossible' (*SC*, 4.8.10). The interesting question here, though, is how to understand the parallels between the strains of commitment and *amour-propre*: what, more precisely, is the role that the strains of commitment play in Rawls' theory, and how is that role illuminated by the parallel with *amour-propre*? What are the demands that meeting the strains of commitment imposes, what does meeting them protect, and how does that compare to the demands made by *amour-propre* and what goods those demands ought to realise? Are societies which fail to meet the strains of commitment as disfigured as societies in which *amour-propre* is corrupt?

The strains of commitment in the broad sense that I am using the term refers to the set of considerations that Rawls argues show the superiority of his two principles to utilitarianism from the point of view of the original position. The reason these considerations matter is because of what Rawls calls 'formal constraints on the concept of right' (*Theory*, p. 130). Those constraints structure the choice that the parties to the original position are offered, since the principles they choose must meet the constraints Rawls lays out. Two are relevant here, finality and publicity. As I read him, Rawls uses the strains of commitment to refer to the difficulties that any given theory has in meeting the requirements of finality, and means by stability a theory's problems relating to the requirement of publicity. However, for my purposes here it is helpful to be able to group the difficulties utilitarianism has with the two constraints under one heading. Since those difficulties relate to the alleged disfiguring of lives lived under utilitarian principles and so the strain that commitment to those principles would mean for some of those living under them, I use the strains of commitment as a portmanteau term to refer to those difficulties.

Publicity requires that principles of right must be 'publicly acknowledged and fully effective moral constitutions of social life': 'general awareness of their universal acceptance should have desirable effects' (*Theory*, p. 133). Finality means that principles of right must be 'the final court of appeal in practical reasoning' (*Theory*, p. 135). Because 'reasoning successfully from [principles of right] is conclusive', '[w]e cannot at the end count [claims of existing social arrangements and self-interest] a second time because we do not like the result' (*Theory*, p. 135).

The thought is that the two principles do well under considerations of finality because they 'insure [the parties] against the worst eventualities' and hence avoid making them party to 'undertaking[s] that in actual circumstances they might not be able to keep' (*Theory*, p. 176). If we could not accept reasoning from principles of right as conclusive, then they would not be final. Since utilitarianism does not ensure parties against the worst eventualities, we cannot be sure that we can accept reasoning from utilitarian principles as final.

The way in which publicity creates difficulties for utilitarianism is slightly more complicated. First, there are considerations of whether a conception of right generates its own support, primarily through 'the psychological law that persons tend to love, cherish, and support whatever affirms their own good' (*Theory*, p. 177). Additionally, there are considerations to do with the public expression of 'men's respect for one another', since '[u]nless we feel that our endeavours are honoured by [others], it is difficult if not impossible for us to maintain the conviction that our ends are worth advancing' (*Theory*, p. 179, p. 178). The two principles do better than a principle of utility under these considerations since 'the principle of utility seems to require a greater identification with the interests of others' (*Theory*, p. 177). They also do better because they include everyone's good 'in the scheme of mutual benefit and this public affirmation in institutions of each man's endeavours supports men's self-esteem' (*Theory*, p. 179). Conversely 'it is natural to experience a loss of self-esteem… when we must accept a lesser prospect of life for the sake of others' (*Theory*, p. 181).

By considering both the reasons for imposing these constraints on principles of right and the reasons for thinking that the two principles do better under them than utilitarianism, we can understand which goods meeting the strains of commitment in my broad sense protects. Both the constraints and the reasons for thinking that the two principles do better under them than utilitarianism seem motivated by a concern with autonomy or agency. The finality constraint could be read as a simple assertion of the primacy of moral reasons. However, given that it is operating as a constraint on the decisions made by the parties to the original position, it seems best interpreted as a kind of responsibility criterion. The parties must bear the consequences of their decision in the original position. The requirement of finality requires that parties take their role as both subject and legislator of justice seriously. They must understand themselves as giving themselves a law. Potential principles they could not treat as the 'final court of appeal in practical reasoning'—because they required excessively onerous sacrifices for others—cannot, Rawls claims, be given to ourselves as laws in this sense.

The publicity constraint, I think, is best read as an attack on false consciousness about the social system one lives within (see for example Freeman 2007a, p. 6).[5] Cultural dupes are not full agents. If I do not understand the normative rationale for the institutions my life is shaped by and within, then I am not exercising full control over that life. It is instead being directed by powers external to me, those which

[5] Andrew Williams' use of a publicity restriction against G. A. Cohen's incentives critique of Rawls defends that restriction on the grounds that it makes possible 'a willing identification with the social constraints to which one is subject' and a 'common pursuit of shared ends' (Williams 1998, p. 244). However, this is not the sense I am interested in.

dictate the shape of those institutions. I cannot guide my action by the correct maxims of conduct, because I do not know what those maxims are. I can hardly be giving myself a law when because of the opacity of the putative law's content, I do not know how to use it to guide my actions. Unless a doctrine justifying coercive political institutions can be publicly proclaimed without undermining those institutions or their justification, then individuals cannot be authors of those institutions or the lives they lead within them. Being systematically misled about the principles dictating the structure of one's life and the institutions one lives it out in is not autonomous.

Both demand that one could live with the results of one's decision, directly in the case of finality and more obliquely in the case of publicity. One only has reason to care about that though, if one has reason to care about one's life in particular rather than some more directly aggregative consideration. Consequentialists of the sort indifferent to the boundaries between individuals, for example, have tended to reject these kinds of constraints (see for example de Lazari-Radek and Singer 2010). More, the way in which utilitarianism fails the constraints are clearly to do with the ways in which one's acceptance of it would disfigure one's ability to live a life. Utilitarianism may require people to live under a social system whose rules that left them in an unacceptable position and which relatedly systematically denied them a sense of their own importance.

Crucially, then, what meeting the strains of commitment responds to the importance of is one's sense of one's own worth, and particularly one's sense of one's own worth relative to and as dependent on others. A bargain to 'acquiesce in a loss of freedom over the course of [one's] life for the sake of a greater good enjoyed by others' 'exceed[s] the capacity of human nature': 'we might wonder whether such an agreement could be made in good faith at all' (*Theory*, p. 176). The claim is that we cannot countenance making that kind of sacrifice because we cannot value ourselves so poorly compared to others. Similarly, the first point about publicity is one about the difficulty of valuing the good of others over our own—'the principle of utility seems to require a greater identification with the interests of others' (*Theory*, p. 177). The second is about the public expression of 'men's respect for one another', and the value that that has in terms of one's ability to pursue one's own projects (*Theory*, p. 179). Having a sense of your own worth in these ways is for Rawls a central feature of being an agent able to direct your own life. To do that, you need a secure sense that you live under a social system which takes your needs and aspirations seriously and which publicly affirms that it does so. Otherwise, at the limit, you would be a cultural dupe whose ability to prescribe yourself laws and hold yourself to them is systematically disrespected and as a consequence likely to have atrophied.

This is not so dissimilar from Rousseau's picture of a society in which *amour-propre* runs wild. There, those on the wrong end of inequality are enslaved to their betters' capricious passions, yet celebrate that slavery: they speak 'contemptuously of those who have not the honour of sharing it' (*Discourse*, p. 187). However, since the masters are ruled by their passions, they are no better off. They are feted, but by those who hate them, and are just as dependent for their sense of self on that obsequiousness as those who offer it are. Just as Rawls notes that '[s]elf-respect is

reciprocally self-supporting' (*Theory*, p. 179), Rousseau does not think that the materially well-off benefit from standing at the right end of inequalities in an unjust society. They are not only vulnerable to all kinds of violence unrestrained by motives of justice or fellow-feeling (see for example *Discourse*, p. 186), but also 'forever active… constantly agoniz[ing] in search of ever more strenuous occupations', desperately seeking either 'a position to live, or renounc[ing] life in order to acquire immortality' (*Discourse*, p. 187).

Members of such a society are not full agents. Widespread conflict means they lack security, but more, they are enmeshed in systems of mutual dependence which deprive them of the ability to properly order their own lives. Because their political system does not guarantee them a stable sense of their own worth, they are dependent on the inconstant and unpredictable valuations of other individuals and so descend into a kind of Hobbesian war of all against all over status. They are in the state of nature again but whilst the first was pure, this one 'is the fruit of an excess of corruption' (*Discourse*, p. 186). Whatever set of terms the social contract of such a society might have, they ask citizens to make agreements 'which exceed the capacity of human nature': they are terms which do not provide them with a secure sense of their own worth, and so terms that agents, who need such a sense, cannot live under.

It is obviously significant here that the development of *amour-propre* is centrally linked to our ability to think of ourselves as agents. As Rousseau puts it, moving from 'the state of nature to the civil state' substitutes 'justice for instinct' in human conduct (*SC*, 1.8.1). If it were not for the fact that 'the abuses of this new condition did not often degrade [citizens] to beneath the condition [they have left], [they] should ceaselessly bless the happy moment which wrested [them] from it forever': their 'faculties are exercised and developed' and their 'ideas enlarged' and 'sentiments ennobled' (*SC*, 1.8.1). They move from being 'a stupid and bounded animal' and become 'an intelligent being' (*SC*, 1.8.1). If humans in the state of nature were little more than beasts, then they, like other beasts, lack the capacity to be agents in the sense that the typical human adult presently possesses. It is reason and the various faculties which make us into moral agents which makes us vulnerable to *amour-propre*'s corruption and the attendant miseries. It therefore, makes sense, on Rousseau's understanding of the evolution of *amour-propre*, to think of it as a demand for respect for one's status as an agent and so parallel to features of Rawls' view which articulate that demand.

Because of this, it also seems a mistake to contrast Rousseau's alleged irrationalism with Rawls' alleged rationalism. The capacities associated with purposive agency are also centrally related to *amour-propre* and the ways in which purposive agency might be undermined or disrespected are also ways in which corrupt *amour-propre* might be inflamed. If *amour-propre* can be satisfied only by a position of common regard and common respect, then it is at least similar to the dispositions to 'love, cherish, and support whatever affirms their own good' and that make it 'difficult if not impossible for us to maintain the conviction that our ends are worth advancing' unless 'we feel that our endeavours are honoured by others'. Rawls talks about dispositions that need to be satisfied in order to meet the constraints on the concept of right, and Rousseau talks about the intellectual

capacities that *amour-propre* is related to and awakens. What meeting the strains of commitment respects and what is disfigured when *amour-propre* is corrupt are not only similar, but have their similarity pointed to by Rawls.

The Value of Constraints on the Concept of Right

The question then becomes, given that similarity, how is Rawls' own view illuminated by the parallel he draws? That the possibility of a just society is constrained even by benign *amour-propre* does not have to constrain what justice is. We might just not be very well suited to justice. Justice might be feasible for angels but not for us. What matters is that we have a reasonable standing interest in satisfying the demands of our sense of self-worth, and crucially must do if justice is to treat us as agents at all. The constraints on the concept of right which generate the problem of meeting the strains of commitment are responses to concerns about agency and autonomy. The reason we care about finality and publicity is that without those constraints on the concept of right, it would treat us as merely its subjects, rather than also its legislators, and hence not as autonomous. Laws which cannot be known cannot be willed, and neither can ones which require us to sacrifice ourselves totally to others. If justice is to be enacted, it is to be enacted through our acts, our agency. If agency deserves any respect then justice will have to take care not to disfigure our agency by making excessive demands on it. Such demands would enslave us to justice.

More, if justice itself aims at protecting agency, it must take care not to disfigure itself by not seeing that its enactment raises questions about its content. A putative demand of justice which in trying to protect or respect agency requires the destruction of someone's ability to coherently direct their life runs the risk of being self-contradictory. Agency can only protect itself by limiting itself, and if those limits are too constrictive, then what is asked for is not protection but self-destruction. If justice is about protecting or respecting agency, then it must not prevent us from acting as agents, and hence must not make demands on us that disfigure our sense of self in the way that corrupt *amour-propre* does. Given the obvious centrality of the two moral powers to Rawls' idea of justice, which together make up an idea of agency, the role of meeting the strains of commitment in protecting the possibility of the exercise of agency must also be crucial to Rawls' theory.

To put it another way, what is illuminated by the parallel is the demand that justice be congruent with our good. What if justice has to be 'grounded in false beliefs covertly instilled in us' or is 'anchored in submission to authority' (Freeman 2007b, p. 152, p. 153)? What if it is a rationalization which serves to 'mask a lack of self-worth and a sense of failure and weakness' or 'a kind of psychological catastrophe… requiring abnegation of the self and its higher capacities' (Freeman 2007b, p. 153)? What if the shared project of justice is one in which distances us from each other, destroying the possibility of other, more straightforwardly communal ends like those of the family or association? If 'a disposition to… justice is destructive of what is best in human character', then it is unclear why we should

care about it (Freeman 2007b, p. 149). Virtues can be debunked, and if justice lacks any resources to explain its importance, then it will be vulnerable to such debunkings. Saying 'because it is justice' presumes rather than demonstrates the presence of such resources. If unjust societies systematically deny us a sense of our own worth, and more, such societies have just some of the features that Rousseau ascribes to *amour-propre* run wild in the *Discourse*, we have such resources. Equally, if justice is an important part of ensuring that we do not face such a society, then not only is it not a threat to our sense of our selves as agents, but it can protect it. Justice needs to be something it makes sense for agents to pursue. If not, it will fail to do proper respect to agency, and so will be vulnerable to debunkings which implicitly rely on the value of the exercises of agency which justice rules out, debunkings much like that which Rousseau provided for the France of his contemporaries.

This is significant because this does not seem to be how these two constraints have generally been understood in the literature. For example, G. A. Cohen's critiques of defences of Rawls against his incentives argument depend on seeing two requirements which together generate what I have called the strains of commitment as independent of justice *qua* justice. Contrasting his view with Rawls', he claims that 'while publicity is indeed a desideratum of rules of social regulation in certain areas, it is surely not a requirement of justice itself' (Cohen 2008, p. 325). We may think that there are various good effects which flow from doing justice publicly, but Cohen wants to separate those good effects from justice itself. They are mere desiderata of rules of social regulation, and maybe not always that (see Cohen 2008, pp. 344–371). He goes on to say similar things about the finality requirement (Cohen 2008, pp. 327–330).

That claim, however, at best operates at a tangent to the kind of explanation of their significance which I have thus far given here. All other things being equal, a putative principle of justice which requires that people are systematically mistaken about the basis of the political institutions they all support and so mutually impose on each other is unsatisfactory *qua* principle of justice, fundamental or not. Similarly, a putative principle of justice which fails to generate its own support because the policies needed to realise it are ones which tend to demean and disfigure the lives of some of those who live under them, is, *ceterus paribus*, an unsatisfactory principle of justice. Indeed, Cohen's recognition of an unspecified personal prerogative which provides an internal constraint on the just pursuit of equality might be thought to be a tacit, untheorised acceptance of such a constraint (see for example Cohen 2008, p. 181). In the absence of a personal prerogative, people would be required to sacrifice their own projects to provide more equality and so be unable to give themselves any special weight in their deliberations in much the same way as is problematic for utilitarianism.

Now, of course, all other things never will be equal. The constraints would be fairly meaningless if all they ever prevented were gratuitous harms, so of course we should expect that they require us to bear costs in terms of what we can get from our principles of justice. What we should do in the face of that, though, is not to deny that these are questions about what justice can ask of us, switching that discussion into one of rules of regulation. Instead we should admit that these issues bear on the

question of what justice, *simpliciter*, is. G. A. Cohen is quite happy to see the question of the permissibility of slavery as bearing on whether or not some principle is a fundamental principle of justice (Cohen 2008, pp. 264–266). The question of the permissibility of 'government house' regimes presumably can play the same sort of role. That is not to say it must play such a role, any more than questions about the permissibility of slavery must. However, a failure to consider seriously the possibility that it might seems unsatisfactory. If it, and with it, questions about one's life lived from the inside, can play that role as a check on the plausibility of some principle of justice, what often amounts to simply ignoring the issue seems both polemically and philosophically unsatisfactory. The proper site of debates about these constraints is instead questions about the costs that would fall under each and whether they ought to be thought of as ones that justice can require us to bear. Perhaps Rawls was wrong about the possible costs of living under utilitarianism and how to assess their importance.

Nothing I have said here rules that out, and indeed the analogy with Rousseau suggests that it could be much more difficult than Rawls seemed to believe to generate conditions under which the constraints could be satisfied. For instance, Frederick Neuhouser has attacked what he calls 'Kantian' readings of what it would take to safely satisfy *amour-propre*. According to him, 'they err insofar as they suppose that the demand to be respected as an abstract person, equal to all others, can or should completely replace the desire to be esteemed for one's particular excellences' (Neuhouser 2008, p. 67). If that is true, then satisfying the Rawlsian constraints I have called the strains of commitment will not tame corrupt *amour-propre* and Rawls is wrong to compare the two. The strains of commitment focus on respecting individuals' agency, not what they have done with it. If esteem for one's particular excellences is necessary to avoid the pathologies of corrupt *amour-propre*, then a Rawlsian society may be riven by them. Pointing to the way Rawls draws attention to similarities between the strains of commitment and the demands of *amour-propre* is not claiming they are identical.

Conclusion

By using what Rawls does with Rousseau in the *Lectures* to outline the motivations underlying the constraint I have called the strains of commitment, I do intend to have ruled some things out. I hope to have shown that it is not adequate to treat considerations of publicity and finality as straightforwardly irrelevant to questions of what counts as a principle of justice. Our intuitions about the justice of principles which require slavery, widespread and systematic deception, or the treatment of some as mere means to ends provided by others seem to indicate that there are some costs which putative principles of justice may not be able to bear. Of course, we may be wrong about that. We may come to revise our beliefs about what costs principle of justice can impose in light of other commitments, either theoretical or practical. G. A. Cohen, for example, was sceptical that justice ought to be 'the final court of appeal in practical reasoning' as a result of his commitment to value pluralism (Cohen 2008, pp. 302–304). Consequentialist appeal to the intuitive

plausibility of impartial maximization as effectively definitive of the moral seems to serve a similar purpose by casting doubt on the plausibility of constraints which might prevent it.

There are two distinct questions here. First, there is the question of what the constraints on the concept of right are. This asks whether appeals to justice really ought to be 'the final court of appeal in practical reasoning' or whether, as Cohen believed, justice is a virtue to be traded off against others in our all-things-considered judgments of what we are required to do. Second, there is the question of how to understand the constraints on the concept of right: whether it matters that we tend to love, cherish and support things which affirm our good, if indeed we do. How much and when should we compromise with human weakness when formulating principles of justice? What are the costs, in terms of our individual plans, projects and purposes, of living under a regime of justice? Cohen, for example, obviously feels that requiring sacrifice of one's own ends to achieving total equality is too much, since he allows for this unspecified personal prerogative to depart from what would otherwise be the demands of justice as a matter of justice. How large should this prerogative be though? Is what it is exercised in defence of it important, or are agents to be left to decide for themselves what gives them an entitlement of justice to ignore justice? Theorising the constraints on the concept of right more completely than has been done thus far would offer a framework in which to consider these questions.

As it stands though, much of the literature—and particularly the literature on Rawls—does not address that task, leaving it missing significant parts of the point and vulnerable to various sorts of debunking. In drawing attention to various potentially fruitful comparisons between Rawls and Rousseau, I hope both to have drawn attention to the work that constraints on the concept of right may do and to have pointed to some of the resources that might be used to theorise them. That, I hope illuminates, at least a little, the question of 'finding a form of association which defends and protects with all common forces the person and goods of each associate, and by means of which each one, while uniting with all, nevertheless obeys only himself and remains as free as before'.

Acknowledgments I would like to acknowledge the generous support of the British Arts and Humanities Research Council who funded the doctorate during which ancestors of this paper were written. It emerged out of a seminar run by Chris Brooke, and I would like to thank him for that and for comments on the paper. Patrick Tomlin also provided comments then, as Christian Schemmel and Chiara Cordelli did later, all of which were very useful. I also owe a debt of gratitude to the referees and editors of *Res Publica* who helped me clarify my claims and their presentation. Finally, it was presented at Manchester University and Nuffield College, and I would like to thank the audiences there, and especially Martin O'Neill and Chandran Kukathas.

References

Cohen, Joshua. 2010. *Rousseau a free community of equals*. Oxford: Oxford University Press.
Cohen, Gerald Allan. 2008. *Rescuing justice and equality*. Cambridge, MA: Harvard University Press.
de Lazari-Radek, Katarzyna, and Peter Singer. 2010. Secrecy in consequentialism: A defence of esoteric morality. *Ratio* XXIII: 34–58.

Dent, Nicholas. 2005. *Rousseau*. London: Routledge.
Frazer, Michael L. 2010. The modest professor: Interpretive charity and interpretive humility in John Rawls's *lectures on the history of political philosophy*. *European Journal of Political Theory* 9: 218–226.
Freeman, Samuel. 1994. Utilitarianism, deontology, and the priority of right. *Philosophy & Public Affairs* 23: 313–349.
Freeman, Samuel. 2007a. The burdens of public justification: constructivism, contractualism, and publicity. *Politics, Philosophy & Economics* 6: 5–43.
Freeman, Samuel. 2007b. *Justice and the social contract: essays on rawlsian political philosophy*. Oxford: Oxford University Press.
Labukt, Ivar. 2009. Rawls on the practicability of utilitarianism. *Politics, Philosophy & Economics* 8: 201–221.
Neuhouser, Frederick. 1993. Freedom, dependence, and the general will. *Philosophical Review* 102: 363–395.
Neuhouser, Frederick. 2008. *Rousseau's theodicy of self-love: Evil, rationality, and the drive for recognition*. Oxford: Oxford University Press.
Rawls, John. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Rawls, John. 1993. *Political liberalism*. New York: Columbia University Press.
Rawls, John. 2001. *Justice as fairness: A restatement*, ed. Erin Kelly. Cambridge, MA: Harvard University Press.
Rawls, John. 2007. *Lectures on the history of political philosophy*. Cambridge, MA: Harvard University Press.
Rousseau, Jean-Jacques. 1997a [1762]. *The social contract and other later political writings*, ed. and (trans: Victor Gourevitch). Cambridge: Cambridge University Press.
Rousseau, Jean-Jacques. 1997b [1750/1755]. *The discourses and other early political writings*, ed. and (trans: Victor Gourevitch). Cambridge: Cambridge University Press.
Williams, Andrew. 1998. Incentives, inequality, and publicity. *Philosophy & Public Affairs* 27: 225–247.