PUBLIC REASON

Journal of Political and Moral Philosophy

Volume 1, Number 2, June 2009

Editura Universității din București

PUBLIC REASON

Journal of Political and Moral Philosophy

Public Reason is a peer-reviewed journal of political and moral philosophy. *Public Reason* publishes articles, book reviews, as well as discussion notes from all the fields of political philosophy and ethics, including political theory, applied ethics, and legal philosophy. The Journal encourages the debate around rationality in politics and ethics in the larger context of the discussion concerning rationality as a philosophical problem. *Public Reason* is committed to a pluralistic approach, promoting interdisciplinary and original perspectives as long as the ideal of critical arguing and clarity is respected. The journal is intended for the international philosophical community, as well as for a broader public interested in political and moral philosophy. It aims to promote philosophical exchanges with a special emphasis on issues in, and discussions on the Eastern European space. *Public Reason* publishes three issues per year in February, June, and November. At least one issue per year is devoted to a particular theme. *Public Reason* is an open access e-journal, but it is also available in print.

Public Reason is available online at **http://publicreason.ro**
ISSN 2065-7285
EISSN 2065-8958
Print version published by **Editura Universității din București** for *Public Reason*.
Comanda Nr. 2465/2009. Tipografia E.U.B.

PUBLIC REASON

Journal of Political and Moral Philosophy

Vol. 1, No. 2, June 2009

ARTICLES

BOOK REVIEWS

BOOK NOTES

Public Reason 1 (2): 1-11

Enlightenment and Constraints

Joseph D. Lewandowski
The University of Central Missouri

Abstract. Drawing on recent work in social philosophy and rational choice theory, in this paper I argue that the core thematic of Kant's "What is Enlightenment?" is the relationship between reason and constraints. I discuss in some detail Kant's definition of and distinction between private and public uses of reason. Most generally, I maintain that while Kant's sense of the private use of reason is too narrowly conceived, his cosmopolitan notion of the public use of reason is far too broad. As a more robust alternative, I propose an account of constitutive constraints and characterize more fully what it means for individuals to make reflexive use of reason vis-à-vis such constraints.

Key words: constraints, Enlightenment, freedom, reason, reflexivity.

The explicit concern of Kant's "What is Enlightenment?" is inarguably that of the power of the public use of reason. Yet there is, or so I want to claim, an even more fundamental question at issue here, namely, that of the relationship between reason and constraints. In fact, throughout "What is Enlightenment?," originally published in the 1784 edition of the *Berlinische Monatsschrift*, reason's role in ascertaining the enabling and limiting conditions of certain kinds of constraints is crucial to Kant's argument. Early on in the text Kant asks: "But which sort of constraint (*Einschränkung*) hinders enlightenment? And which, instead of hindering it, can in fact enhance it?" (55).[1] Kant's well-known answer is that while tightly constraining the private use of reason enables civil order and a government's procurement of "public ends" (56), the public use of reason must always be unconstrained, and "it alone can bring enlightenment to the human condition" (55). In pursing this line of inquiry it is perhaps not surprising that Kant concludes his reflections by noting the way in which a "lesser degree of civil freedom" can actually ensure ever-greater degrees of "intellectual freedom" and, over time, the opportunity and capacity to "act freely" (59).

The issue Kant wrestles with here, commonly known as the paradox of choice, is that less is often times, but not always, more. In fact, to put it somewhat crudely, I think that the challenge laid out in "What is Enlightenment?" is to use reason to determine when less is more and when it is not. Or, to describe the matter in terms to be elaborated here, the critical task of enlightenment in Kant's sense is to make reflexive use of reason to optimize constraints. With the phrase "reflexive use of reason" I mean simply the embedded and embodied capacity of human beings to make explicit and alter the conditions that enable and limit thought and action. By "optimize constraints" I mean modifying and/or creating the kinds of rules and norms that maximize human freedoms of thought and action. Of course I shall endeavor to clarify and provide examples of what is meant by

1] Translation modified, as are all subsequent citations of the English edition (1970) of Kant's "What is Enlightenment?" cited here.

these terms in what follows. But to state my thesis briefly: the purpose of the reflexive use of reason is to get the various constraints of society right. Indeed, if the ultimate objective of enlightenment is freedom, as Kant suggests, then the proper motto of enlightenment is, "optimize constraints!"[2]

Now, the contemporary philosophical landscape offers several different paths for re-thinking Kant's concerns in "What is Enlightenment?". But two paths in particular provide a useful contrast for framing the argument to be developed here. The first, blazed by Nietzsche but extensively widened by Nietzschean-inspired "post"-modernists, is pursued in a decidedly skeptical and un-Kantian way. Here appeals to the use of reason, however conceptualized and operationalized in various historical moments, are viewed as inherently masking relations of domination and the will to power. History, politics, science, knowledge, morality – all these are construed as rationalizing processes of human subjugation in which contingent constraints of the dominant eventually harden into structures and systems that establish the un-free order of things. Consequently, in the words of Foucault, it is only in an "historical ontology of ourselves" (1984, 45) that the enlightenment's critical engagement with constraints can be redeemed. In such a "post"-modern ontology what is sought is an "historical analysis of the limits that are imposed on us and an experiment with the possibility of going beyond them" (1984, 50).[3] In place of the use of reason, this path of existential experimentation, where Nietzsche and Foucault are joined by Bataille (1985; 1993), Heidegger (1991), and Derrida (1978; 1985), seeks an *aesthetic rapport a soi* where internalized limits may be transgressed in what Foucault once cryptically described as "techniques of management" (1983, 18) of the self.

By contrast, there is a well-established second route that does not entail a skeptical point of departure from Kant. Instead, it seeks to provide a positive account of the conditions needed to realize the project of enlightenment in a fundamentally Kantian way. In fact, this approach, developed most prominently by Rawls (1971; 1999) and Habermas (1991), aims to reconstruct the emergence of an actually existing "public sphere" and defend a normative account of the use of reason in such a sphere. With their demands that individuals detach themselves from the social milieu, substantive identities, and historical experiences that shape them, the contemporary heirs to Kant's formulation of the public use of reason aim to create the necessary conditions for democratic inclusion. Hence for neo-Kantians such as Rawls and Habermas, rational deliberation constitutes the normative medium of the public use of reason, while social abstraction remains the precondition for the empirical realization of such reasoning in actual public space.

In my discussion here I should like to consider the relationship between reason and constraints from a somewhat different angle. Rather than draw on the current work in neo-Kantian philosophy or pursue strong "post"-modern critiques of that line of thinking,

2] For a related treatment of Kant, see Brandom 1979.

3] For a critical analysis of Foucault's account of power and agency, see especially Habermas 1987; 1989.

I shall introduce recent work in social philosophy and rational choice theory to outline an account of constitutive constraints, and then connect that account to the reflexive use of reason aimed at constraint optimization. The overarching purpose of pursuing such a discussion is two-fold. On the one hand, I want to clarify and characterize more fully the constitutive nature and function of constraints in Kant's thinking. On the other hand, I want to propose an orientation toward those constraints that is neither over-determined by relations of power nor under-determined by abstract appeals to what Kant calls a "public of world-readers" (55). On my account, enlightened agents are best understood as highly reflexive constraint-optimizers who are no more reducible to effects of power than they are inflatable to free-floating members of a world public. Along with a gain in conceptual clarity, the virtue of such an account is that it engages Kant's thinking about enlightenment in a way that does not require a theoretical description or empirical realization of "the public."

Consequently it must be emphasized at the outset that the goal here is not to rehabilitate a neo-Kantian account of the public use of reason. Nor, however, is it my intention to contribute to the general skepticism that clouds the prospect of the use of public reason in contemporary life.[4] Rather, my principle aim is to give more precise definition to a conception of constraints that is implicit but under-developed in Kant's text, and then to characterize what it means to adopt a reflexive orientation toward those constraints. My argument, in sum, is that it is not publicity but rather an account of reflexivity that is decisive for scrutinizing the relationship between reason and constraints. Indeed, as I hope to show, the "public" use of reason in Kant's sense is best understood as a distinctly reflexive use of reason.

Let me begin, then, with a brief summary of Kant's attempt to distinguish sharply between private and public uses of reason. Two examples offered in "What is Enlightenment?," although not altogether analogous, as we shall see, are particularly illustrative of the importance of the relationship between reason and constraints in Kant's thinking here. The first is that of the private use of reason deployed by a military officer. As a member of a military organization, such an individual finds himself situated in a rigidly constrained matrix of chain-of-command type rules and codes of conduct. In this context, as Kant argues, "it would be very harmful if an officer receiving an order from his superiors were to quibble openly, while on duty, about the appropriateness or usefulness of the order in question. He must simply obey"(56).

Yet the private use of reason is crucial not simply to maintain obedience and order. Its use is also essential because it is precisely in the ongoing acceptance of and adherence to the shared constraints (rules and codes) of a military's organizational scheme that certain individuals can be defined and count as officers. For being an officer consists in thinking and acting (i.e., taking and executing orders) in strict accordance with the jointly shared

4] But for two insightful discussions of the deficits of current theories of public reason, see especially Fraser 1997 and, more recently, Hrubec 2008.

rules and codes that define and make possible a military organization. In the absence of such constraints, it is not only difficult to see how an individual could be considered an officer but also how a military could exist at all. It is precisely shared enabling constraints (rules and conduct codes) that, at least at one basic level, *constitute* a military.

Kant's example of the clergyman should be similarly construed. The clergy, too, finds himself in a context rather narrowly defined and yet enabled by a jointly shared set of constraints – though those constraints are perhaps better thought of as associational beliefs and norms rather than rules and codes, as in the case of the military. In the course of his daily labors, the clergyman is bound by the enabling constraints that define and make possible his position and the religious group to which he belongs. Accordingly for Kant, he must make private use of reason in his work. For what it means to count as a clergy within a particular religious association is primarily to represent and disseminate the established beliefs and norms that define that association. In the presence of his congregation, therefore, the clergyman is, as Kant says, obliged to say: "Our church teaches this or that, and these are the arguments it uses" (56). In other words, for Kant the narrow task of the clergy qua clergy, like that of the officer qua officer, is to embody and express the defining-enabling constraints of the association, and not to make explicit or call into question those constraints from within the narrow confines of what Kant characterizes as the "purely private" space of a "domestic gathering" (57).

In his account of the private use of reason in military and religious contexts Kant has thus identified the way in which less can indeed be more – the way, that is to say, that certain sorts of constraints define and enable the existence of societal organizations and associations. Yet Kant also realizes that while necessary, such a use of reason vis-à-vis enabling constraints is not sufficient in an enlightened society. For while individuals must accept and adhere to the constraints of the various organizations and associations in which they are embedded – as rule- and code-followers or belief- and norm-applicators – they nevertheless require a standpoint from which to address emergent situations when less is *not* more. What is needed, in other words, is an orientation from which to *criticize* constraints when they become sub-optimal and no longer enable in ways that they could, should or were designed to do. In his discussion of the clergy Kant stresses the need to allow for such a critical orientation when he insists that "it is absolutely impermissible to agree, even for a single lifetime, to a permanent religious constitution which no one might publicly question" (58). Now for Kant, as we know, such "public questioning" must be disembedded from societal constraints, wherein only the "private" use of reason is allowed. Indeed, Kant formulates the genuinely "public" (*öffentliche*) use of reason precisely as a "free" or an unconstrained way in which individuals may orient themselves toward and gain critical purchase on various organizational or associational constraints. In making public use of their reason, officers and clergy are thought to be able to socially unbind themselves and enter into a "real public" (57) as "men of learning" (56), "scholars" (57) and "world-citizens" (56) limited only by a "rational respect for personal value and for the duty of all men to think for themselves" (55). Thus, the officers and clergy shed their "private"

identities and speak in their "own person" (56) as "scholar[s] addressing the real public" (57) regarding "the errors in military service" or the "better arrangement of religious and ecclesiastical affairs" (56) . In short, when more is or has become less, Kant proposes the public use of reason.

There are four inter-related issues raised by Kant's distinction between the uses of reason and sense of constraints that I should like to consider here. To begin with, it must be pointed out that Kant's use of the term "private use" (*Privatgebrauch*) is conceptually confusing. Clearly what he means – and what all of his examples illustrate – is something decidedly *un-private*. Military organizations and religious associations – as well as bureaucracies and the various venues of civil society inhabited by Kant's other example of tax-paying citizens – are *social* sites. By "social sites" what is meant here is simply those collective locations or contexts where the actions, beliefs, attitudes and identities of individuals are in various ways interlocking and interdependent – where, that is to say, a sense of "we" creates "plural subjects," to borrow Margaret Gilbert's (1989) useful conception. Military organizations and religious associations are plural subject phenomena insofar as they are comprised of individuals whose individual mental states *necessarily* include the shared consciousness of a unity and commitment to undertake joint actions *with others*. The conceptual point to be clarified is that individual officers and clergy don't merely make "private" use of reason in the respective contexts of their daily work; rather, they count as officers and clergy precisely because they must individually reason in ways that reflexively adopt and incorporate the constraints shared by *other individuals* who exist within their organizational or associational "we." Hence the "private" use of reason in Kant's sense is really one of the *reflexive* uses of reason vis-à-vis societal constraints – I shall return to this point below.

Second, in "What is Enlightenment?" societal constraints have, as we have seen, a distinct function to which Kant alludes but does not adequately develop. Specifically, the function of such constraints is to define and enable: that is, they create and maintain the possibility of certain shared ways of thinking and acting. In other words, societal constraints are *constitutive* constraints. Unlike hard constraints (such as that of gravity or technical limits such as those that once limited film-making to silent movies), constitutive constraints are those soft bounds that both constitute and are intentionally constituted by certain plural subject entities.[5]

The definition of constitutive constraints I want to articulate here may also be understood by way of John Searle's distinction between regulative and constitutive rules.[6] In his work on the construction of social reality, Searle (1995) argues that:

5] In an extended discussion of Durkheim and Gilbert, I have sought to explain how constraints – understood as "social facts" – can be both objectively given to and subjectively made by human actors. See Lewandowski 2002.

6] A related but much earlier discussion of this distinction already appears in Rawls 1955.

> Some rules regulate antecedently existing activities. For example, the rule "drive on the right-hand side of the road" regulates driving; but driving can exist prior to the existence of that rule. However, some rules do not merely regulate, they also create the very possibility of certain activities. Thus the rules of chess do not regulate an antecedently existing activity… Rather, the rules of chess create the very possibility of playing chess. The rules are *constitutive* of chess in the sense that playing chess is constituted in part by acting in accord with the rules. (27-28)

What I have been calling "societal constraints" (or what in Kant's examples might more accurately be designated organizational and associational constraints) are constitutive in much the same way that Searle's constitutive rules are: in both cases such constituting limits do not merely regulate but also create what counts as playing chess or being an officer or clergy. Indeed, as we have seen, what Kant misleadingly calls the "private" use of reason can exist only within a system of shared constitutive constraints.

Third, the existence and ongoing maintenance of constitutive constraints takes place in contested fields of thought and action, and thus the so-called private *use* of reason is much more sociologically complex than Kant's account suggests. Put simply, where societal constraints are present, less is almost always more for some, and not for others.[7] Or, to put the point in Searlean terms: while the constitutive rules of chess can be said to be equally enabling and constraining for all players, the same cannot be said for the constitutive constraints of societal organizations and associations. In fact, while constitutive societal constraints are by definition jointly shared, they can and often do function in profoundly stratifying ways to create what Pierre Bourdieu (1977) calls a "habitus."[8] That is to say that such constraints characteristically engender positions of privilege among some of those who share them, as well as conditions of exclusion for many of those outside of them. In the constitutive constraints of many military organizations and religious associations throughout the world, for example, women simply cannot count as officers or clergy.

Moreover, those individuals who are able to achieve positions of ascendancy within exclusionary military organizations and religious associations do so at least in part because of their more or less successful attempts to navigate and gain control over the various material and symbolic goods available within those constraints. Indeed, among other factors, becoming an officer or clergy involves acquiring a feel for maneuvering within and out-maneuvering others in struggles for power within particular organizational and associational constraints. It is precisely in this way that, for example, the sense of "we" of a

7] Numerous everyday examples spring to mind. To elaborate just one: it is a safe bet that most international airline travelers would prefer one pre-determined gourmet meal to the prevailing in-flight choice between chicken and vegetable pasta. But if I am member of a vegan culture and the pre-selected meal is, say, steak au poivre, then clearly less is not more for me in this case.

8] Specifically, Bourdieu defines habitus as "the durably installed generative principle of regulated improvisations" (1977, 78). I have elsewhere discussed the relative strengths and weaknesses of Bourdieu's account of habitus and theory of practice (Lewandowski 2000).

military organization constituted largely by hierarchical command chains often becomes so stratified that the officers' sense of "we" does not include the men they command.

Finally, Kant appears to over-reach in his characterization of the "public" use of reason as a kind of cosmopolitanism outside of *all* constitutive constraints. To return for a moment to the example of the military officer alluded to above: imagine that such an officer wanted to address the problem of stratification within his military organization. As already noted, while some degree of organizational stratification may be necessary for the functioning of a military, too much stratification is clearly sub-optimal, and threatens the kind of plural subjecthood required for the successful execution of operations in the field. On Kant's account, an officer who sensed that existing constraints no longer sufficiently enabled is enjoined to adopt an orientation entirely outside of the constitutive constraints that define him as an officer or member of a nation-state and speak simply as a "man of learning" to a ''world public."

Yet while it is obvious that an officer should be free to reason in a way other than what Kant mistakenly calls "private," appealing to a specifically "public" use of reason is unwarranted. Indeed, it is difficult to see how – or why – a military officer would shed the many layers of his social skin, as it were, and address something like a "world-public." The point to be made is not that officers cannot or should not reason from a perspective other than that of members of a particular branch of the military or specific nation-state. Rather, what must be admitted is that their rational dialog and critique will inevitably be informed by the constitutive constraints that create the conditions of possibility of their identities, roles and experiences as actual "men of learning." That is to say that officers are officers; clergy are clergy. And their "publics" are what the social has made them out to be, as Kant himself ambiguously acknowledges when he says that an officer cannot be banned from submitting his judgments about errors in the military to "*his* public" (*seinem Publikum*) (my emphasis, 56).[9]

Of course officers are not only officers, and clergy are not only clergy. Inasmuch as they are also enlightened agents, they are all reflexive participants *and* observers in the various organizations, associations and diverse life-worlds they inhabit. But it is precisely for this reason that distinctly "private" and "public" uses of reason find no place in such agents' ways of reasoning. The choice enlightenment presents is not between making private or public use of reason, but rather among various reflexive uses of reason vis-à-vis shared constraints, as I shall try to make explicit below.

In short, the reflexive relationship between reason and constraints is not adequately elaborated in "What is Enlightenment?" The sources of this shortcoming should by now be apparent: Kant misleadingly speaks of a "private" use of reason, places too heavy an emphasis on the definition of and distinction between the private and public uses of rea-

9] To complicate matters further: what constitutes a military man's "public" is not at all obvious. For example, in his study of WW II American and German soldiers, Shils (1951) demonstrates that it is primarily the constitutive constraints of small groups that define and enable the "we" of an effective military.

son, and does not sufficiently characterize the constitutive nature of societal constraints. Nevertheless, in "What is Enlightenment?" Kant does rightly emphasize the importance of scrutinizing constitutive constraints. Yet it is not in transcending such constraints in the name of publicity but rather in the reflexive use of reason vis-à-vis constraints that Kant's claims about enlightenment are best understood.

In the previous portions of this paper I have thus sought to clarify the constitutive nature of such constraints, and to highlight some of the ways in which constitutive constraints are enabling and limiting. In the remainder of my discussion I should like to say a bit more about what is meant by the term "reflexivity."[10] As I understand and use it, the term reflexivity characterizes the relationship between reason and constraints implied in both of Kant's uses of reason. The use of reason is reflexive to the extent that it seeks to optimize constitutive constraints at various moments and in various ways. Such reflexivity can be paradigmatically found in one of three forms: choice of constraints; interrogation of constraints; and the creation of new constraints.

In its most basic form, a reflexive orientation towards constraints can be found in the everyday exercise of rational choice. As we have seen, in one fundamental sense officers and clergy are simply those who have elected to adhere to one set of shared constraints rather than another. Of course it hardly needs to be said that human actors' choices, and the paths available to realize those choices, are never unlimited. As Jon Elster argues, all human choices are the result of two successive filtering devices:

> The first is defined by the set of structural constraints which cuts down the set of abstractly possible courses of action and reduces it to a vastly smaller subset of feasible actions. The second filtering process is the mechanism that singles out which member of the feasible set shall be realized. (1984, 113)

Or, to put the argument in the terms used here, while all everyday rational choices vis-à-vis constitutive constraints are pre-filtered by "structural constraints," there is nevertheless a kind of cognitive feedback mechanism that monitors and informs which of the available constitutive constraints is to be adopted at various times and in various contexts. On my account, that mechanism is reflexivity. In this way reflexivity complicates any simple or straightforward causal assumption about how the pre-filtering effect of structural constraints might determine individual choice of constitutive constraints. Initially, structural pre-constraints merely reduce the relative range of possible choices human actors may reflexively opt to pursue. Indeed, despite the structural pre-constraints that have narrowed their options, rational actors can and do reflexively orient (and continuously re-orient) themselves towards the constitutive constraints that remain open to them as they seek to realize their changing preferences and goals in diverse contexts. Thus, for example, while poverty may be a rather severe structural pre-constraint on an individual's feasible set of constitutive constraints, it is not causally determining of a single human

10] My discussion of reflexivity here is in part informed by Bogdan 2000 and Bourdieu & Wacquant 1992. But see also Lewandowski 2000.

choice, future course of action, or expression of values: a poor man's choice of religion or armed service is not in any necessary way a mere expression of his economic limitations.[11]

Additionally, and to paraphrase Elster (2000), along with the everyday reflexive choice of constraints comes a distinct kind of interrogative reflexivity within constraints. Kant, as I have argued, mischaracterizes these two reflexive uses of rational choice as distinctly private and public, or, even more problematically, as the difference between being a "cog in a machine" (56) or a cosmopolitan member of the "world at large" (57). Pace Kant, however, it is not a question of mechanistic obedience or a view from nowhere. Instead, in interrogative reflexivity, agents can and do scrutinize their constitutive constraints in the course of their existence as socially embedded reasoners. For as elective rational participants in the constitutive constraints that define their actions and identities, they are also always already observers. At certain times and in certain places, such participant-observer reflexivity takes mental note – or actively minds – the sub-optimal nature and effects of a given set of constitutive constraints.[12] While at other times and in other places the reflexive use of reason thematizes and makes explicit those sub-optimal elements for others to see.

In both cases, however, it is as participants in and observers of existing constitutive constraints that agents adopt an interrogative stance and communicate their rational critiques to other individuals. As we know, this latter use of reflexivity is what Kant calls the public use of reason. Yet my argument here is that in such cases what is entailed is not the use of reason outside of constitutive constraints but rather the reflexive use of reason with regard to such constraints. Indeed, when less is not more, it is interrogative reflexivity that calls into question sub-optimal bounds. It is precisely in this way that reflexive participant-observer critiques of sub-optimal constraints can and do lead to the transformation of existing constraints. Critical discussions about military hierarchies, for example, can foster related conversations about the larger function and purpose of such organizations (and perhaps war in general), or of the military's sub-optimal use of labor in its exclusion of women, ban on homosexuals, and so on. In sum, optimizing the organizational constraints of a military or the associational constraints of a religious order is inevitably dependent upon the extent to which individuals make reflexive use of reason vis-à-vis such constraints.

Now to be sure, constitutive constraints are not simply rationally chosen or interrogated. There are also unique periods and contexts of human thought and action when entirely new sets of constitutive constraints must be fashioned. In fact, in the present

11] Nor is a limit on his monetary resources causally determining of his everyday choice of something even as basic as transportation. A poor man in Detroit with only two US dollars in his pocket may not be able to afford a taxi and might therefore appear to be structurally preconstrained to travel by public transportation to meet a friend. But that outcome is not pre-determined in any singular way. One can imagine that if this man is in relatively good health he may reasonably opt to keep the money to buy food or clothing and walk to his destination.

12] See Bogdan 2000.

context one needs look no farther than the revolutions and transitions to post-socialism undergone in the preceding decades in Central Europe to see this kind of creative reflexivity at work. For while it might rightly be said that revolutions aim to destroy existing constitutive constraints, successful transitions typically demand the creation of new societal constraints.

Indeed, creating democratic institutions and market economies, however complex and contested, is at its core a constitutive constraint-making endeavor or series of endeavors.[13] The goal of such a highly innovative undertaking is to create what Elster (2000) calls an "optimal tightness of bounds": markets and democracies must be constrained enough to enable efficiency and fairness, yet loose enough to ensure a maximum amount of liberty and innovation. In creating market-based democracies, the reflexive use of reason aims to design and engender conditions in which continued reflexive orientations towards constitutive constraints are possible.

Let me conclude with a brief summary of my position. Kant's "What is Enlightenment?" stands as an attempt to consider the relationship between reason and constraints, and, moreover, as an argument about how such a relationship should be construed in an enlightened society. Ultimately, as we have seen, Kant's characterization of the relationship between reason and constraints depends upon a core distinction between "private" and "public" uses of reason. But as I have argued, such a distinction is both misleading and unwarranted. On my account, both uses of reason outlined in "What is Enlightenment?" should be understood as entailing reflexive orientations vis-à-vis constitutive constraints: where "private" reason entails reflexive choice of and provisional adherence to constitutive constraints, "public" reason involves the reflexive choice within constraints to adopt an interrogative stance when those constraints become sub-optimal. In this way my position shares with Kant the central insight that when less is not more, taking up a critical orientation with regard to societal constraints is imperative for enlightenment. Yet such a critical participant-observer orientation, I have maintained, cannot be located in the ether of cosmopolitanism. On the contrary, it is in the reflexive use of reason vis-à-vis constitutive constraints – and not in the public use of reason beyond such constraints – that enlightenment resides.

lewandowski@ucmo.edu

REFERENCES

Bataille, Georges. 1985. *Visions of Excess: Selected Writings 1927-1939*, trans. Allan Stoekl. Minneapolis: University of Minnesota Press.

———. 1993. *The Accursed Share: Vols 2 & 3*, trans. Robert Hurley. New York: Zone Books.

Bogdan, Radu. 2000. *Minding Minds: Evolving a Reflexive Mind by Interpreting Others*. Cambridge, MA: MIT Press.

13] As evident in Elster et al. 1998.

Bourdieu, Pierre. 1977. *Outline of a Theory of Practice*, trans. R. Nice. Cambridge: Cambridge University Press.

Bourdieu, Pierre, and Loic Wacquant. 1992. *An Invitation to Reflexive Sociology*. Chicago: University of Chicago Press.

Brandom, Robert. 1979. Freedom and Constraint by Norms. *American Philosophical Quarterly* 16 (3): 187-196.

Derrida, Jacques. 1978. *Writing and Difference*, trans. Alan Bass. Chicago: University of Chicago Press.

———. 1985. *Margins of Philosophy*. Trans. Alan Bass. Chicago: University of Chicago Press. Elster, Jon. 1984. *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

———. 2000. *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints*. Cambridge: Cambridge University Press.

Elster, Jon, Claus Offe, and Ulrich Preuss. 1998. *Institutional Design in Post-Communist Countries: Rebuilding the Ship at Sea*. Cambridge: Cambridge University Press.

Foucault, Michel. 1983. Afterword. In *Beyond Structuralism and Hermeneutics*, Hubert Dreyfus and Paul Rabinow. Chicago: University of Chicago Press.

———. 1984. What is Enlightenment?. In *The Foucault Reader*, ed. Paul Rabinow. New York: Pantheon.

Fraser, Nancy. 1997. Rethinking the Public Sphere. A Contribution to the Critique of Actually Existing Democracy. In *Justice Interruptus: Critical Reflections on the 'Post-socialist Condition*, Nancy Fraser. London: Routledge.

Gilbert, Margaret. 1989. *On Social Facts*. Princeton, NJ: Princeton University Press. Habermas, Jürgen. 1987. *The Philosophical Discourse of Modernity: Twelve Lectures*, trans. Frederick Lawrence. Cambridge, MA: MIT Press.

———. 1989. *The New Conservatism: Cultural Criticism and the Historians' Debate*, trans. Shierry Weber Nicholson. Cambridge, MA: MIT Press.

———. 1991. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Trans. Thomas Burger. Cambridge, MA: MIT Press.

Heidegger, Martin. 1991. *Poetry, Language, Thought*. Trans. Albert Hofstadter. New York: Harper Collins.

Hrubec, Marek. 2008. On Conditions of Participation: The Deficits of Public Reason. *Human Affairs* 18: 81-91.

Kant, Immanuel. 1970. An Answer to the Question: 'What is Enlightenment?'. In *Kant: Political Writings*, ed. H. S. Reiss. Cambridge: Cambridge University Press.

Lewandowski, Joseph. 2000. Thematizing Embeddedness: Reflexive Sociology as Interpretation. *Philosophy of the Social Sciences* 30 (1): 49-66.

———. 2002. What Makes a Fact Social? On the Embeddedness of Social Action. *Existentia: An International Journal of Philosophy* 12 (3-4): 281-293.

Rawls, John. 1955. Two Concepts of Rules. *Philosophical Review* 64: 3-32.

———. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press

———. 1999. *The Law of Peoples; with the Idea of Public Reason Revisited*. Cambridge, MA: Harvard University Press.

Searle, John. 1995. *The Construction of Social Reality*. New York: Free Press.

Shils, Edward. 1951. The Study of the Primary Group. In *The Policy Sciences*, ed. Daniel Lerner and Harold D.Laswell. Stanford: Stanford University Press.

Public Reason 1 (2): 12-30

Rawls: Construction and Justification

Stefan Bird-Pollan
Harvard University

Abstract. I examine Rawls' indebtedness to Kant in *A Theory of Justice*, Kantian Constructivism and in "Themes from Kant's Philosophy". I argue that the way Rawls develop the justification of *A Theory of Justice* relies heavily on Kant's claims that rationality requires reciprocity and that rationality is to be understood as moral rather than as instrumental. Rawls thus reveals something new in Kant's theory namely that for Kant the hypothetical imperative is actually subordinate to the categorical imperative. However, Rawls eschews Kant's attempt at proving that we are rational and thus committed to treating each other with respect, hence Rawls argument fails to show that we do, in fact, share the intuitions about justice as fairness that underlie Rawls' theory.

Key words: Rawls, Kant, morality, constructivism, justification.

The metaphysical problems that plagued Kant's deduction of morality in the *Groundwork* III have seemed, to many twentieth century philosophers who wanted to retain much of Kant's moral philosophy, so great that these contemporary thinkers have abandoned the attempt to ground pure practical reason altogether. The question I mean to pursue in this paper is whether a certain type of Kantian moral philosophy can get by without such a grounding. In Rawls one finds a writer who believes that much of Kant's ethical theory can be salvaged if one sidesteps the question of a metaphysical justification for morality and concentrates on the proceduralism necessary for justice.

The question, to put it another way, is whether the Kantian framework that Rawls adopts, lends itself to a non-metaphysical use. By this I mean that Kant's system may be metaphysical through and through and as such require the discharging of certain assumptions in its final form. This ultimate metaphysical assumption, I will argue, is that there is, in fact, not just a shared but a universal morality. The aim of this paper is thus to reconstruct the parallels between Rawls' argument and Kant's own, drawing out just how heavily Rawls leans on Kant to construct his theory. With this parallel in place, it will then be possible to determine whether, given the strong parallels I argue exist, Rawls' theory can still claim to be valid without working through the metaphysical assumptions Rawls explicitly rejects.[1]

Rawlsian constructivism is, as I hope to show, a worthy successor to Kant in the sense that it seeks to avoid the problems that have plagued generations of Kant interpreters – to find some way of making the categorical imperative 'work'. Rawls' strategy, by contrast, is to concentrate on the categorical imperative as a way of thinking about moral laws immanently, that is, as constantly articulated and enacted by the individual agent. For

1] It is interesting to note that several of Rawls' students have returned to the path of a metaphysics of sorts in order to ground the universality of morality. See, for instance, O'Neill 1996, 194, Herman 1993, 198 and in particular Korsgaard, 1996, 15; 2009, 189.

Rawls, the categorical imperative is just the mental process we engage in when we think about how to be just to other human beings. Rawls thus emphasizes respect for persons over moral psychology. Respect for persons entails that we treat others just as we want to be treated by others and this simply means, not seeking special treatment for oneself. Respect, Rawls argues, should (and generally does) enters into every thought about others. This type of thinking is modeled in both of Rawls' justifications for the liberal political society: the original position and the reflective equilibrium. The categorical imperative is a way of thinking which enables such respect for others.

I will argue, however, that, compelling though Rawls' interpretation of Kant's ethical theory is, its aim of presenting a non-metaphysical interpretation is only partially successful. Rawls is successful in giving a non-metaphysical account of reflection through the reflective equilibrium – a process in which each agent reflects on her considered beliefs and also takes into account the beliefs of others. Absent a universal (and therefore 'metaphysical') notion of practical reason which underlies such reflection, however, there is no way of showing that the conclusions of individual reflection cohere in any socially meaningful way. Indeed, this absence of cohesion is the result of Rawls' failure to take concrete suffering into account. By building his theory on the possibility of coherence between individuals, Rawls has, I will argue, sidestepped the problem of the perspective of justice altogether.

A further way of framing the issue is to see Rawls' and Kant's theories as objections to the egoist who believes that all she is committed to is taking the means to her ends, but not to anything further. In believing this, the egoist essentially resists the idea that the hypothetical imperative is framed by the categorical imperative or that the rational is framed by the reasonable. To refute the egoist one must, however, make precisely this move. And this move relies on the metaphysical assumption of our membership in a community which shares the same fundamental commitment to universal justice.

In reconstructing Rawls' thought, I will present the argument regressively, starting from Rawls' conception of autonomy and working backwards, always asking for a justification for the previous level of argument, until at last we arrive at the reflective equilibrium which is supposed to underwrite the whole conception of justice. The regressive reconstruction follows the argument Rawls gives in "Kantian Constructivism in Moral Theory", if not in *A Theory of Justice*, and underlines the acknowledged debt Rawls owes to Kant. The regressive argument also affirms that, after all, Rawls wishes to give a Kantian style grounding to his project since the regressive argument is itself a device used by Kant in order to arrive at a transcendental argument, and argument, I will argue, Rawls fails to deliver.

I. THE RATIONAL AND THE CATEGORICAL IMPERATIVE

In the interest of space, I will not spend much time on Rawls' twin concepts, the original position and the veil of ignorance. They both model what Rawls will call the 'ra-

tional' in "Kantian Constructivism". Suffice it to say that for Rawls, the original position is a regulative principle and thus a way of adjudicating between conflicting desires and inclinations.[2] The agent in the original position must be both autonomous and motivated by her reflection. She takes the means to her ends. That is to say, the original position must yield universally acceptable principles (as in the hypothetical imperative which, for Kant, is analytic) and it must ensure that these principles are acceptable to all.[3] The former condition is modeled in the original position by bargaining and the latter is modeled by the veil of ignorance.

Let us look at autonomy first. Rawls introduces the veil of ignorance to hide the parties' particular social and natural circumstances. The parties are asked to design a society without the knowledge about where they will be placed in the society, or which beliefs, moral, political or religious they will have.[4] All participants understand the basics of political affairs and economics and possess general knowledge. Thus they choose principles under which they are prepared to live, wherever they end up in society. The general social structure is just but blind to the particular inclinations of the agents. Under the veil of ignorance, just as in Kantian autonomy, we have no personal or particular sense of the good. We seek only justice, the ability to enjoy our particular notion of the good once we determine what that is.

Rawls also argues that there is a parallel between rational choice theory and the categorical imperative. Rawls says that the original position is in the tradition of social contract theory. Like the categorical imperative, it provides a way of responding to a practical problem: what ought I do? Rawls' two principles of justice are simply the moral law under the conditions of a modern liberal society, yielding more specific versions of the universal prescriptive of respect as stated in the categorical imperative.

We should note two points before we go on. In the model of the original position, Rawls has moved moral reflection from the first person perspective to public deliberation; from the 'I' to the 'we'. At least prima facie, the original position is not supposed to be all in the mind of one individual. The second point follows from the first. By changing the perspective of reflection from the first person to the third person, Rawls has also changed the moral psychology involved in accepting the outcome of deliberation. It is not clear that accepting the outcome of public deliberation has the same normative force as accepting the outcome of my own deliberation on the authority of the moral law.[5]

2] Rawls himself does not believe that Kant's categorical imperative actually provides a particularly good way of determining a content of the moral law. This is what his own theory of justice is supposed to provide (2007, 31).

3] Many have argued that a hypothetical agreement does not constitute a justification for the two principles chosen in the original position. See Nagel 1975, 114.

4] There has been considerable objection to the supposed neutrality made possible through the veil of ignorance. Onora O'Neill, for instance, notes that Rawls does not assume disinterest at all times during the original position process, but permits it with reference to the fate of future generations (1998, 121).

5] In a way, this is the problem Rawls will have to address in *Political Liberalism* where he will have to

Indeed, the central argument for the universality of morality hangs on this move from subjectively accepted norms to universally accepted norms. How is it, one might ask, that norms I develop for myself in my interactions with the world should be acceptable to all others? To put it another way, what is Rawls' argument against the egoist who believes that reasons are essentially private. Rawls seeks to address this issue in what follows.

II. THE REASONABLE AND THE RATIONAL

Rawls agues that underlying the original position and the application of the categorical imperative there is a conception of the moral character of the actors who reflect and thus abide by the moral law. In "Themes in Kant's Moral Philosophy" Rawls interprets these agents as both 'reasonable and rational'. Rawls uses these terms as a translation for Kant's *vernünftig*, which includes both senses. The two terms mark the distinction Kant makes between the two types of practical reason, pure and empirical practical reason. The former is found in the categorical imperative while the latter is exemplified by the hypothetical imperative. Rawls notes that Kant's conception of a person also marks the fact that, for him, the hypothetical imperative (empirical practical reason) is absolutely subjugated by the categorical imperative (pure practical reason) (1999a, 112). This is to say that the person who engages in moral reflection subjugates his rationally conceived maxims to the moral law.

Rawls characterizes his project in "Kantian Constructivism" as the attempt to: "establish a suitable connection between a particular conception of the person and the first principle of justice, by means of the procedure of construction" (37). This means that Rawls attempts to construct a philosophically coherent story about how the idealized conception of the person as reasonable and rational, can lead to a set of public institutions of justice we all can endorse. Before we examine what Rawls means by constructivism, we must understand what he means more exactly by the reasonable and the rational.

In political terms this means:

> [W]henever a sufficient basis for agreement among citizens is not presently known, or recognized, the task to justify a conception of justice becomes: how can people settle on a conception of justice, to serve this social role [of admissible social institutions], that is (most) reasonable for them in virtue of how they conceive of their persons and construe the general features of social cooperation among persons so regarded? (1999b, 305)

To put the issue slightly differently than Rawls does, we could say that the hypothetical imperatives each person at the bargaining table wishes to realize are limited by the recognition that each of the bargainers is equal and that it is thus unreasonable for one member to insist that the group agree to make an exception for that member. Thus the reasonable which models the demands of universality in the categorical imperative

---

show that we accept the results of the original position for reasons that are in a sense pure or moral rather than prudential. For a classic formulation of the objection to this move see Williams 1985, 205, ch. 4.

frames the debate about which particular hypothetical imperatives can be realized. The notion of universality, which Rawls interprets as equality, frames and restricts the particular rational plan of any actor. This turns classical liberal 'negative' freedom into a more communal 'positive' freedom. Thus when Rawls says that the original position is morally neutral, he means that there is no conception of the good involved in decision making itself. Morality, however, *is* in play in the sense that freedom and equality have a particular moral perspective, which is that the reasonable frames the rational.

But in order for the rationality of the original position to yield more than prudential agreement or a modus vivendi, Rawls must show that having a thin theory of the good allows each agent to move to a thick theory of the good. This is the point of introducing the distinction between the rational and the reasonable.[6] Rawls wants to show that instrumental reason as employed in the original position can be seen as an ethical capacity from a different perspective. This leads to a reinterpretation of the original position in "Kantian Constructivism", which relies more heavily on the notion of equality than its predecessor in *A Theory of Justice* did.

The movement from third person perspective to first person perspective occurs in three stages. It starts from rational autonomy (bargaining proper), moving to full autonomy (bargaining with reasonable or moral constraints) and finally ending up with the readers of Rawls' theory themselves (which finds its justification in the reflective equilibrium). What Rawls calls the rational or rational autonomy is modeled in pure procedural justice.[7] At the second stage, of full autonomy, Rawls adds to the conception of the person as free and equal two moral powers and two higher-order interests. The first power is that of having an effective sense of justice, the second is the power to form and revise and rationally purse a conception of the good. Corresponding to these are the higher-order interests of realizing and exercising these powers (1999b, 312).

The move to full autonomy and the reasonable, Rawls writes, is "expressed by the framework of constraints within which the deliberations of the parties (as rationally autonomous agents of construction) takes place" (1999b, 317). This framework is the reasonable ideal of fair cooperation. The framework, by which Rawls means the addition of the two moral conceptions of the person, reciprocity and mutuality, ensure that the plan of the good each person articulates for him or herself also includes the good of others. This is the doctrine of respect for persons as it is expressed in Kant's second formulation of the categorical imperative, the formula of humanity.[8] Here people are conceived of as an ends in themselves. Thus the two moral powers overlay the process of rational deliberation, transforming the instrumental deliberative process in the original position into a process of mutual recognition and fair cooperation. Rawls elaborates: "In justice as fairness, the

6] See Rawls 1993a, 503-4, and 1999b, 316. Also Baynes 1992, 122.

7] This means that the outcome is justified by if the means of arriving at it were just.

8] "Act so that you use humanity in your own person or in the person of any other, always at the same time as an end, never merely as a means." (Kant 1996, 58, Ak 4:429)

Reasonable frames the Rational and is derived from a conception of moral persons as free and equal. Once this is understood, the constraints of the original position are no longer external." (1999b, 319). I take this to mean that only the device of the original position (which models instrumental reason) imposes the constraint of fair cooperation on the people. For the people in the original position, social cooperation is not intuitive. But it is so for fully autonomous people who live in the institutions which the two principles of justice have helped to create. For they see themselves as possessing the two moral powers and thus restrict their pursuit of the good in the name of something more than the maximization of their material gain.

The movement of the two stages so far trades on the distinction between different perspectives. If we move back a little, we might recall that the purpose of the original position is to develop principles of justice out of our presuppositions about moral character. That is, what kind of laws would free and equal people come up with if left to their own devices? What Rawls does is to draw out first what free or rational individuals would do and then to overlay this with what people who are both free and reasonable would do. Rational people seek to maximize their benefit while reasonable people seek to maximize their benefit with the concerns of others in mind. This parallels exactly the structure that Kant argues for as well: we are rational beings insofar as we try to realize our ends by adopting the means to do so, but we are moral insofar as we adopt only those ends which we can will others to adopt as well.

Thus Rawls can say: "The unity of practical reason is expressed by defining the Reasonable to frame the Rational and to subordinate it absolutely; that is, the principles of justice that are agreed to are lexically prior to their application in a well-ordered society to claims of the good." (1999b, 319).

The lexical ordering of the reasonable over the rational also parallels Kant's division of practical reason into empirical practical reason and pure practical reason. While empirical practical reason– the hypothetical imperative– means acting according to any practical principle, pure practical reason– the categorical imperative– means acting according to the principle of the moral law.

However, there are still two elements missing from this argument. The first, to which we will now turn, is the question of how we get from the presupposed character of the agent as reasonable and rational to the content of the principle of justice, which so far has been described only formally. The second question, which we will come to after that, is what justifies the assumption of people as 'reasonable and rational' in the sense of being free to set their own goals. The second question comes down to what grounds Rawls assumption that we are, in fact, reasonable (or moral) and hence that I set *my* goals with other people's goals in mind.

III. CONSTRUCTIVISM

Constructivism is meant to be the way to get from a certain conception of the person (here, free and equal) to the appropriate principles of action for such a person. This means that constructivism seeks to draw out the content of the conception of the agent and to formalize it. That is, if the CI-procedure is the appropriate form of a rational principle, what is the appropriate material? The answer is the free and equal agent.[9] It is the answer to the question: what should we do when we act under the moral law or use our pure practical reason (which amounts to the same thing)?

In Kantian terms this means that: "the totality of particular categorical imperatives (. . .) that pass the test of the CI-procedure are seen as constructed by a procedure of construction worked through by *rational* agents subject to various *reasonable* constraints." (Rawls 1999c, 513-14). Each time we reflect and determine a law for ourselves we construct an element in a universal set of rules which can then be abstracted and turned into a general duty. Rawls' two principles of justice are a version of what might be arrived at in such an abstraction. The point, though, is that the maxims of conduct permitted or enjoined by rational reflection are not theoretical speculations; they are responses to actual needs for clarification of the permissibility of intended action.[10]

We can thus say that constructivism is the idea that the content of our highest moral principles stems from the rational and reasonable reflection upon our concepts as free and equal agents. Constructivism models autonomy in the sense that it constitutes the moral law or principle of justice from within its own rational and reasonable reflection. Nothing can count as a law for me without my having determined it for myself. This strongly echoes Kant's claim that there is nothing good in itself except the good will.[11]

Now, as before, there is here an emphasis on the first person perspective. That is, constructivism is just the CI-procedure insofar as it pertains to determining the content of the moral law. The content of the moral law has the content it has because I have (rationally) reflected upon it and have determined that it has this content. We must, however, keep open the possibility that when this first person perspective is switched to a third person perspective, as it is in rational choice, we loose normativity altogether. We will return to this issue.

Construction thus has two elements. First, it is a process internal to the agent and as such it is from a first person perspective. No one can reflect for me. Second, it is practical. Since reflection on the permissibility of performing an action stems from an incentive for action, the result of my reflection can only ever be manifested in my action itself. The

9] For this way of putting the problem see Korsgaard 1996, 123.

10] O'Neill notes that the constructivist position is anti-realist because it denies that moral facts are discoverable in theoretical terms. Constructivists believe that ethical principles are constructed by human agents, that these principles are practical and that they are objective. See O'Neill 2003,. and also Korsgaard 1996, 124.

11] See *Groundwork*, Ak 4:393.

result of my reflection can only ever be what I actually do, that is, what motivates me. If I say I ought to give $100 to charity and do not, I have actually decided to keep the $100. A practical constructivism thus relies on a notion of pure practical reason, that is, the idea that we are capable of reflecting on our ends by the use of the moral law or the two principles of justice.

Let us take a step back again and see where the argument has gotten us so far. Constructivism was introduced to provide a link between the conception of persons as free and equal (or as reasonable and rational) and the content of the principles of justice. Rawls' contention was that through the process of construction, or through autonomous reflection, these ideal agents would determine a set of principles which are able to govern the agents who have developed them in fair cooperation. Construction was then the way to bring out the content of the basic idea of the reasonable and rational agent without introducing any alien conceptions of how the world is or ought to be. The only tool available to the reasonable and rational agent in determining what the principles of justice are is reason. We also found that this constructivism proceeds from a practical point of view, which cannot be justified in theoretical terms. Justice is immanently constituted as doing that to which all involved have agreed.

Thus three of the four elements of Rawls' argument are in place. The original position has been established as yielding a universal principle. The presuppositions about the moral character (freedom and equality, reasonable and rational) of the agents who participate in the original position have been examined. And lastly, constructivism has presented a way for us to move from these presuppositions of moral character to the actual content of the formal characteristics of the moral law: the principles of justice.

The only element that is still missing is the justification of why we should think that we are actually those people in the original position who frame the rational by the reasonable. That is, what makes me think that I can assume that other people share the concern I have them. Where, in other worlds, does the universality or the reciprocity of the reasonable and rational framework lie. This is, to be clear, the basic assumption about morality that Kant was unable to provide in the *Groundwork* and an assumption which Rawls must make good on if the original postion, the reasonable as framing the ration and constructivism are to make sense.

IV. EXCURSUS: RAWLS AND KANT, PARALLEL ARGUMENTS

I have argued that Rawls has roughly followed the structure of the *Groundwork*. He has developed the categorical imperative in terms of the universality of the two principles of justice, corresponding to the first formulation of the categorical imperative.[12] Then Rawls has switched perspectives and has argued that the presupposition for such a law

12] "I ought never to act in such a way that I could also will that my maxim should become a universal law." (*Groundwork*, Ak 4:402)

is that people respect their humanity and, finally, Rawls has contended that in order to act under the moral law, we must imagine ourselves as instantiating the two principles of justice. For Kant this is the kingdom of ends. Let us examine these parallels in a little more detail.

Rawls writes: "In the first formulation [of the categorical imperative], which is the strict method, we look at our maxim from our point of view. (. . .) We are to regard ourselves as subject to the moral law and we want to know what it requires of us." (1999c, 505). This, I want to argue, is similar to the original position in which we want to know the formal structure a principle of justice would have.In the second formulation, however, we are to consider our maxim from the point of view of our humanity as the fundamental element in our person demanding our respect, or from the point of view of other persons who will be affected by our actions. Humanity both in ourselves and in others is regarded as *passive*: as that which will be affected by what we do (1999c, 505).

As I have already indicated above, I take this to be the perspective of drawing out the presuppositions about agents in the original position. To frame the rational by the reasonable means to see ourselves as passive in the face of the hypothetical imperative and to try to avoid damage to our humanity by restricting its scope. Our humanity is the material for the application of the CI-procedure in the sense that this is the purpose for that procedure. Rawls adds: "The point is simply that all persons affected [by my will] must apply [the CI-procedure] in the same way both to accept and to reject the same maxims. This ensures a universal agreement which prepares the way for the third formulation." (1999c, 505)

"In [the third] formulation we come back again to the agent's point of view, but this time we no longer regard ourselves as someone who is subject to the moral law but as someone who makes the law. The CI-procedure is seen as the procedure adherence to which, with a full grasp of its meaning, enables us to regard ourselves as legislators– as those who make universal public law for a possible moral community." (1999c, 506) This last formulation is clearly analogous to constructivism in the sense that in constructivism we develop positive law out of our conception of ourselves as free and equal.

I provide this juxtaposition of the structure of Rawls' and Kant's arguments not only to support Rawls' claim that *A Theory of Justice* is largely Kantian in orientation but to show that *A Theory of Justice* brings out central features of constructivism which must be seen as not just incidental but substantive contributions to Kant scholarship (§40). I further wish to argue that by tying his theory to Kant so closely, Rawls' theory is subject to many of the same difficulties as Kant's work. These difficulties have mainly to do with the problem of justification. For instance, the failure of Kant's deduction of morality has left Kant without a footing from which to say that humans are indeed able to interact respectfully with one another. Because Rawls avoids this push toward immanence and stays at what might be called the 'common sense' level, he also lacks a philosophically rigorous conception of intersubjectivity. Rawls' rejection of metaphysics, as I have said before, leaves him without an answer to the question of how people can actually be relied upon to treat each other with respect.

V. JUSTIFICATION AND THE REFLECTIVE EQUILIBRIUM

If the theory of construction is the justification for the two principles of justice, then what justifies construction? Rawls' answer, like Kant's answer to the problem of why humans should consider themselves free, is quite simply that constructivism is not justified in a theoretical way, but is given its authentication through cohesion into the perspective of existing humans who find that they agree with it. Justification is given through action. This notion of coherence is the final step in the three part development of authentication presented in "Kantian Constructivism".

Finally, Rawls comes to consider the last perspective, "that of ourselves– you and me – who are examining justice as fairness as a basis for a conception of justice that may yield a suitable understanding of freedom and equality" (for our own practical use) (320-21). Rawls continues:

> Here [in the third perspective] the test is that of general and wide reflective equilibrium, that is, how well the view as a whole meshes with and articulates our more firm considered convictions. (. . .) A doctrine that meets this criterion is the doctrine that, so far as we can now ascertain, is the most reasonable for us. (1999b, 321)

At this third perspective then, we have arrived at the criterion for a final justification of Rawls' theory. The problem for Rawls, as for Kant, is that we cannot prove that people believe themselves to be those ideal agents. Rawls is quite convinced that the failure of Kant's deduction of the moral law is sufficient to show that such an idealized theory approach makes no sense. So, according to the third perspective, the justification or authentication comes down to what Rawls calls the reflective equilibrium.

Let us now examine what the reflective equilibrium is in more detail. As Kenneth Baynes puts it: "reflective equilibrium refers to a condition in which an individual's concrete moral judgments have been brought into harmony with her higher-order moral principles" (1992, 69). This harmonization occurs first through a narrow process of reflective equilibrium in which one moves back and forth between concrete judgments (in, say, the manner of the categorical imperative in which the subject decides on a maxim and, using the categorical imperative procedure, determines whether it can be acted upon – if not, a new maxim must be created and tested) and then through the wide process of reflective equilibrium in which one's own judgments are bought into harmony with general social norms, shared by most readers of *A Theory of Justice*.[13]

Let us now turn to Rawls' own characterization of the process before turning to criticisms and defenses of this method. Rawls holds that his theory of justice describes our own sense of justice (1999a, 35). The justifications of his theory of justice, modeled by the original position and background conditions, are all reflections of our own considered

13] Rawls already characterizes the discussion about Justice as Fairness as taking place within a bounded society, one that endorses liberal democracy. Readers coming from outside this realm, may not agree with him.

judgments. The need to write *A Theory of Justice* in the first place, however, must have been generated by the knowledge that, on the face of it, not everyone currently *does* in fact share Rawls' conception of justice. The task of justifying the theory of justice thus must occur though a process of fleshing out those beliefs we actually all hold.[14]

The process of achieving reflective equilibrium systematizes our beliefs.[15] What we do in the narrow reflective equilibrium is thus similar to what we do in the CI-procedure. We take a practical problem which, admittedly, is more abstract than our everyday practical concerns, and reflect on it. For what is systematizing but bringing disparate concepts under a general principle of practical reason.

There is thus a positive and a normative side to the reflective equilibrium, or, as Thomas Scanlon put it, a descriptive and a deliberative side. As a method of arriving at an accurate portrait of justice, we must dig within ourselves to find normative notions we endorse (2003, 113). Both sides seem to be included in the following statement by Rawls: "we do not understand our sense of justice until we know in some systematic way covering a wide range of cases what these principles are." (1999a, 41). Indeed, in this statement of the purpose of the reflective equilibrium it is not possible to separate the two senses. Since, however, the process of the reflective equilibrium is a theoretical undertaking to which we subject our considered judgments, it seems appropriate to call it a method of deliberation.

The method itself is not explained in great detail in *A Theory of Justice*.[16] I will thus cite only the two main passages from this work in which Rawls describes the reflective equilibrium process as it pertains to original position:

> By going back and forth, sometimes altering the condition of the contractual circumstances [in the original position], at others withdrawing our judgments and conforming them to principles, I assume that eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted. (18)

> A conception of justice cannot be deduced from self-evident premises or conditions on principle; instead, its justification is a method of the mutual support of many considerations, of everything fitting together into one coherent view. (19)

The reflective equilibrium begins with our considered opinions which must be made "under conditions favorable to deliberation and judgment in general". (40) A cen-

14] Who exactly the 'we' is, has been the subject of much debate. See, for instance, Okrin 1994, 125

15] See Rawls' formulation about the original position: "The conditions embodied in the description of the original position are ones that we do in fact accept. Or, if we do not, then perhaps we can be persuaded to do so by philosophical reflection." (1999a, 19)

16] Rawls' idea of the reflective equilibrium has been taken up in the fields of moral philosophy and the philosophy of science. See, for instance, the more rigorous formulation James Blanchowicz gives (which is not based strictly on Rawls' account). Likening the reflective equilibrium process to repairing a ship at sea, Blanchowicz writes: "It is not just the fact that one is resting on a dry part of the ship in one's efforts to repair a leaking part and that one may later rest on the repaired (formerly leaking) part to repair a new leaking (formerly dry) part that establishes genuine reflective equilibrium, but rather the fact that the ***way*** in which one rests on these respective parts is different in each case (. . .)." (1997, 126)

tral feature of these conditions, and one which will be relevant to criticism of the reflective equilibrium discussed below, is that the process of the reflective equilibrium is always subject-dependent. That is, it is always my judgment that comes in for consideration. In this sense, judgments are judgments only when they come with my reasons for the judgments attached.[17] They are thus not comparable to observational data.[18]

As a coherentist strategy, convergence in reflective equilibrium is only evidence of how much agreement we already have. It is not normative, in the sense that it might convince one who does not hold what I hold to change his or her mind. It is purely introspective. This is because, as Norman Daniels argues, coherentism in the form of the reflective equilibrium remains agnostic about whether there is any truth which it might approximate. This agrees with the point about anti-realism raised earlier according to which constructivism develops all 'truths' through reflection itself.

VI. PROBLEMS WITH THE REFLECTIVE EQUILIBRIUM

As we saw above, the method of the reflective equilibrium is a way of becoming clear about one's own ethical convictions. We examine our thoughts and our principles are measured against what we might have read or discussed with others. The important point to keep in mind is that we are now ourselves, comprehensive subjects with commitments to notions of the good. This means that each of us reflects from a different perspective.[19]

At this point we must, however, make a distinction which I mentioned earlier, namely the distinction between the content of the theory and its very possibility. For there is an ambiguity in the charge that we reflect from different perspectives as real existing agents. The charge might mean that, since we are different, we are not sure whether we will come to the same conclusions as Rawls does. But it might also mean that we would have completely different conceptions of morality or that morality might be denied altogether. The former point is addressed by the bulk of Rawls' argument while the latter point refers to a problem Rawls does not have much to say about.[20]

17] This is perhaps the place to note that Rawls never develops an adequate justification of the reflective equilibrium from the first person perspective, and thus ultimately leaves himself open to criticism from Kantians and others who regard the subject as the primary unity of ethical coherence. See the discussion of Christine Korsgaard and Onora O'Neill below. Both seek to remedy this deficiency in Rawls' account through their respective theories of practical reason.

18] See Daniels 1979, 12: 167-72. See also Brandt 1990, 128

19] This is what Sandel has in mind when he argues that the Rawlsian deontological subject is incapable of normative commitments because he or she has been cleansed of all contingency which would necessitate normativity in the form of judgment. In order to make the Rawlsian subject capable of normativity, normativity must be introduced at a later stage but this is impossible given the thinness of the subject as it is conceived in the original position (Sandel 1982).

20] Rawls addresses this issue in *Political Liberalism* where he talks about commitment to the liberal state as opposed to the modus vivendi, a temporary commitment which, in certain extreme cases, might seeks to overthrow the whole system.

The problem of different starting points for reflection brings with it a host of problems for Rawls. For instance, there is no longer any compelling connection between the perspectives of the different subjects being asked to endorse Justice as Fairness. The point is put nicely by Baynes who argues that there seems to be no reason for me to accept the results of the reflective equilibrium unless I am the one who has undergone the process myself. This, presumably, is what Rawls means when he writes that "each person has in himself the whole form of a moral conception" (1999a, 44). Here, whole must mean complete for me and not, as in Kant, universal. Thus, there does not seem to be any reason why I should be swayed by a subjective process of reasoning not my own.[21] We are thus back at the question of who the 'we' who endorses the considered moral judgments is and whether there is any connection among the individuals which make up the 'we'. In *A Theory of Justice* and in "Kantian Constructivism", this 'we' seems not to have been theorized at all where possible justification is concerned. This leaves open the possibility of egoism and hence the possibility that we do not, in fact, deliberate together as Rawls believes we do.

Scanlon, similar to Baynes, argues that the reflective equilibrium process is normatively underdetermined. This charge states simply that no conclusive evidence for or against Rawls' theory can be gotten from a coherentist justification.[22] Since the reflective equilibrium does not offer a determinate process by which one might arrive at ethical conclusions, it is quite possible for two people to start from the same premises and, using the reflective equilibrium method, still arrive at different conclusions. Rawls acknowledges this point when he says that his theory of justice is just 'a' theory of justice (1999a, 43-44).[23] But as a theory of justice it must include the claim that something is normative for us even if we cannot agree entirely on what it is. The problem is thus that a coherentist theory which seeks its justification in the reflective equilibrium is too weak to bind people of differing perspectives together because it cannot on its own overcome the differences that people with previous normative commitments bring to bear on their reflections. Coherentism, in other words, seems not to be able to provide consensus where there is none to begin with.

There is another, deeper objection here, however. Scanlon has argued that someone's employment of the reflective equilibrium commits the evaluator of the argument who undertakes it to nothing at all.[24] This question delves deeper since it asks the more

21] See Baynes 1992, 74.

22] Brandt argues, for instance, that Rawls' argument comes to a conclusion no more forceful than that: "A coherent set of beliefs can be made more convincing than another set even if there is nothing which can confirm or refute it." (1990, 272-73)

23] Concerning the intersubjectivity of the reflective equilibrium process, Rawls writes that the question must remain open: "I shall not even ask whether the principles that characterize one person's considered judgments are the same as those that characterize another's. I shall take for granted that these principles are either approximately the same for persons whose judgments are in reflective equilibrium, or if not, that their judgments divide a few main lines represented by the family of traditional [moral] doctrines (…)." . Rawls adds, referring to himself, that: "if we can characterize one (educated) person's sense of justice, we might have a good beginning toward a theory of justice". (1999a, 44)

24] See Scanlon 2003, 152 and O'Neill 1998b, 206-7.

fundamental question of whether morality exists at all and thus lays bare the assumption Rawls has so far been making about the reflective equilibrium, namely that it *is* the pure employment of practical reason. If the reflective equilibrium is, in fact, the pure employment of practical reason, there will be no problem with coherence beyond the merely technical problem of the correct assessment of the facts. We thus need some further argument about why the reflective equilibrium is, in fact, the employment of pure practical reason and not some other principle. This goes to the more fundamental question of the possibility of morality and thus relates quite clearly to Kant's own failed attempt at proving intersubjectivity.

The problem I am here insisting on is that the answer to the problem of the justification of the two principles of justice in the reflective equilibrium cannot be gotten through an analysis of the coherence of the two principles of justice with our own perspective as readers of political theory through the reflective equilibrium. The deeper problem suggested here turns on the question about the possibility of morality in general, which cannot be answered by coherentism precisely because it is a question of first principles or metaphysics, if you will. Indeed, coherentism can only give an evaluation of the rightness or justice of two principles of justice if it is assumed that coherence is really an expression of morality or practical reason.

Before taking up this final issue, we must look a little more closely at what the role of pure practical reason is in Rawls' theory. And this crucially depends on the perspective employed in the philosophical reasoning of *A Theory of Justice*.

VII. PURE PRACTICAL REASON AND THE FIRST PERSON PERSPECTIVE

We have now seen all four elements of Rawls' theory so I now want to take stock of the argument as a whole and make good on the promises for elaboration I made during the reconstruction of the argument. I will thus discuss what I see as the real problem in Rawls' ultimate justification of his theory, by which I mean the position of pure practical reason. In both the original position and the reflective equilibrium Rawls presents us with a conception of normativity, through bargaining and the interpretation of social norms, which seems to want to sidestep the question of the need for a justification of his claim for our ability to employ pure practical reason. I will argue, however, that a notion of pure practical reason must underlie both conceptions. I will then return to the issue of whether pure practical reason receives a foundation in Rawls' work.

The first problem I mentioned was the problem of what I argued was the substitution of the original position for the categorical imperative in *A Theory of Justice*. I noted that in this move Rawls replaced a first person perspective with a third person perspective. He seemed to be arguing that the process of deliberation under the veil of ignorance was just as good at leading to the two principles of justice as solitary reflection. Indeed, the substitution rather suggests that Rawls thinks rational choice is a better model for ethical thought than solitary reflection.

From a Kantian perspective, however, this move seems highly suspect. For what gives rise to normativity in Kant is that I make the law for myself, that I am an autonomous actor. As is clear from the reading I gave above, Rawls also considers this to be the case with the agents in the original position behind the veil of ignorance. But rational reflection, as Kant sees it, operates only from the first person perspective. That is, something is normative for me because I choose to adopt it as a principle. No one else can make me adopt as my end something I do not freely choose as an end. You may force me to do it, but it will not be my end.

This is just the familiar point that practical reason cannot be given a theoretical explanation. No one can convince me by argument that I should adopt their reasons. I must convince myself. So, if deliberation in the original position is really a 'compromise' as Rawls states, then the agreement reached in it is not normative for anyone since it does not represent a principle anyone actually endorses. The principle that has arisen through the compromise might, of course, still be adopted, but Rawls has not given us any argument for why those in the original position should adopt the principles they have reached in negotiation (1999a, 104-5).

There is a way out of this argument, of course. It is essentially that the original position with its multiple parties is just a way of representing what goes on in rational reflection in the CI-procedure. The move from the CI-procedure to the original position is just heuristic.[25] That this is so becomes quite clear, I think, when one examines the notion of constructivism, which is meant to connect the two principles of justice to the original position. Constructivism seeks to draw out the consequences of our presuppositions about the agents negotiating in the original position. But in order for us to be able to draw out anything about them, we must assume that they have something in common, namely the concepts of freedom and equality. This is why Rawls refers to these agents as idealized. In order for the process of construction to yield anything at all, 'idealized' must mean that they are at least generally the same. If this is so, then the move from free and equal individuals through construction to the two principles of justice merely mirrors Kant's movement from the *vernünftig* individual though rational reflection to the moral law.

As such, it is no mystery that the agents in the original position can come to a 'compromise' which is normative for all. The compromise is no compromise, it is really the presupposition of the moral theory underlying the make up of the agents – justice as fairness. There has thus been no shift from the first person perspective which admits of the use of practical reason to the third person perspective. There is also thus no issue of convincing anyone of the rightness of the two principles of justice.

So, as I think I have shown, the problem of normativity of the two principles of justice does not arise at the level of the original position since, fundamentally, the original position models the use of pure practical reason by an autonomous self. This does not mean, however, that the problem of normativity has been laid to rest. The normativity of the two

25] See Dworkin 1975, 129

principles of justice is simply moved back to the reflective equilibrium and to the question of its acceptance by comprehensive subjects. It also does not mean that the problem of the first person to third person switch and the problems this entails has gone away.

According to the arguments I have just given, we must conclude that Rawls' attempt to build greater stability for his system, through both the notion of bargaining in the original position and through the idea of wide reflective equilibrium which ultimately rely on a notion of public reason, is really reducible to the employment of pure practical reason by each individual. And the existence of pure practical reason in finite human beings is, of course, what Kant was unable to show in the deduction in *Groundwork* III.

VIII. THE PROBLEM OF NORMATIVITY AND THE NECESSITY OF ITS JUSTIFICATION

The fact that I have agued that the social anchoring that Rawls wants to give his theory by embedding it in broad social views is inconsistent does not mean, however, that the theory must be rejected or even that its steps are incoherent. I have merely shown that Rawls actually sticks far closer to Kant's general argumentation than is usually supposed. Two theoretical moves have been rejected but as long as we interpret these moves as merely heuristic, the general theory remains intact. It is thus time to come to the question of the final justification of the reflective equilibrium, in other words, whether there is an account of pure practical reason in Rawls' theory.

And here we come to the central problem of Rawls' justification. Kant saw his theory as hinging on the proof or authentication of the necessity of freedom and morality both in the deduction and in the fact of reason doctrine. Rawls does not think his theory requires such a grounding.

This brings us again to the problem of the first person and third person perspective of practical reason. I argued first that the agents in the original position, as autonomous and idealized, must share the same conception of freedom and equality, and that this means that they are really not substantially distinct in a way that would necessitate a compromise in determining the two principles of justice. Then I argued that the reflective equilibrium process which we must all engage in, in order to determine whether we actually believe ourselves to be similar enough to the idealized agents in the original position to endorse the two principles of justice they determine, also had to stem from a first-person reflection. Thus the claim and its authentication both stem from a first-person perspective.

Without a proof for the necessary identity between the results of the original position and the results of our own reflection, the most that can be said of the two principles of justice is that they cohere. And this is all Rawls wants to say. Rawls refuses Kant's deduction of morality in favor of Kant's fact of reason. And Rawls interprets the fact of reason as a coherentist justification for the two principles of justice.

Thus Rawls writes:

> Pure practical reason is authenticated finally by assuming primacy over speculative reason and by cohering into, and what is more, by *completing* the construction of reason as one unified body of principles: this makes reason self-authenticating as a whole. (1999c, 523)

The idea here is that since there can be no theoretical proof of freedom and morality, the only justification for morality that can be given is that we recognize ourselves as moral beings, that is, we recognize ourselves as the agents who participate in the original position.[26] This means, as we saw, that the speculative part of his theory, the original position and the mutual regard of rational autonomy, cannot be justified except through empirical endorsement by fully autonomous actors, the people engaged in ethical reflection – you and me.

Thus, much rides on the notion of recognizing ourselves in the idealized agents of the original position. This recognition comes down to believing ourselves to be capable of employing pure practical reason. And this belief is what Rawls means by cohering into a much broader notion of reason. Rawls thus offers a minimalist authentication of the possibility of morality itself.

I mentioned earlier that Rawls thought that the categorical imperative offered only modest help in determining contentful principles of justice. We are now in a position to make better sense of this claim. Because of his coherentist justification of the two principles of justice, Rawls does not maintain that *his* two principles of justice are the only ones possible. That is, he does not maintain that he has determined the precise content of the laws our social organization should take. He has proposed 'a theory' of justice which is open to revision.

But this is not a significant departure from Kant, since Kant was not proposing a significant content to the moral law. He was only interested in showing that there is such a thing as the moral law. This, however, is a position in which Rawls must follow Kant, since in order for there to be any kind of theory of justice, the possibility of a theory of justice must be given. And this is what Kant's deduction and later his doctrine of the fact of reason seeks to show.

The second claim is deeper, for it contains a thesis about the ultimate justification of morality. Here Rawls just assumes that the principle of practical reason really exists. In this sense, the revisability of the two principles of justice which, as we saw, are supposed to be derivatives or incarnations of Kant's categorical imperative, depends on there being such a thing as the categorical imperative or freedom in the first place. By seeking to give a weaker interpretation of the categorical imperative in terms of the CI-procedure, Rawls has given up on Kant's claim that the weak autonomy thesis must be turned into a strong autonomy thesis. Rawls has, in other words, given up on the idea of showing that humans are rational beings and has just assumed that we are.

26] See also O'Neill 2003, 356-57.

But giving up on the strong autonomy thesis means that, as I have argued, there is no answer to the question raised by Kant, namely, why should we think that we appetitive humans are motivated by rational laws and hence, why should we think that what you think is 'rational' is not just a way of subjugating me. This is one possible objection the egoist might make against Rawls. The whole question of justice, in other words, rests on showing that we are all in possession of a common rationality which can help us to overcome our appetitive natures and adhere to derivatives of the categorical imperative as Rawls or anyone else proposes them. If the possibility of rational agency is not circumscribed by reasonable agency, then the egoist will not be refuted.

To put the question one last way in terms of Kant's analytic and synthetic distinction: it might be analytically true that humans would follow something like Rawls' two principle of justice if they were rational, but to show this goes beyond the scope of Rawls' book. I hope, however, to have raised the issue of the grounding of reason with sufficient urgency to show that metaphysical neutrality is not an option for a theory of ethics.

pollan@fas.harvard.edu

REFERENCES

Baynes, K. 1992. *The Normative Grounds of Social Criticism: Kant, Rawls, and Habermas.* Albany, State University of New York Press.

Blanchowicz, J. 1997. Reciprocal Justification in Science and Morality. *Synthesis 110*: 447-68.

Brandt, R. 1990. The Science of Man and Wide Reflective Equilibrium. *Ethics 100*: 71-90.

Daniels, N. 1979. Wide Reflective Equilibrium and Theory Acceptance in Ethics. *The Journal of Philosophy 76*: 256-82.

Dworkin, R. 1975. The Original Position. In *Reading Rawls. Critical Studies on Rawls' Theory of Justice,* ed. N. Daniels. Oxford: Blackwell.

Herman, B. 1993. *The Practice of Moral Judgment.* Cambridge, Massachusetts: Harvard University Press.

Kant, I. 1996. *Groundwork of The Metaphysics of Morals.* In *Practical philosophy.* Cambridge: Cambridge University Press

Korsgaard, C. M. 1996. Kant's Formula of Humanity. In *Creating the kingdom of ends.* Cambridge: Cambridge University Press.

———. 1996. Reasons We Can Share. *Creating the kingdom of ends.* Cambridge: Cambridge University Press.

Nagel, T. 1975. Rawls on Justice. In *Reading Rawls. Critical Studies on Rawls' Theory of Justice,* ed. N. Daniels. Oxford: Blackwell.

O'Neill, O. 1996. Towards Justice and Virtue. Cambridge University Press.

———. 1998a. The Method of A Theory of Justice. In *John Rawls. Eine Theorie der Gerechtigkeit*, ed. O. Höffe. Berlin: Akademie Verlag.

———. 1998b. Kantian Constructivism in Ethics. *Ethics 99* (4): 752–70.

———. 2003. Constructivism in Rawls and Kant. In *The Cambridge Companion to Rawls*, ed. S. Freeman. Cambridge: Cambridge University Press.

Okrin, S. M. 1994. Political Liberalism, Justice and Gender. *Ethics 105*: 23-43.

Rawls, J. 1993a. Themes in Kantian Moral Philosophy. In *Kant and Political Philosophy*, eds. Ronald Beiner and William James Booth. Yale University Press.
———. 1993b. *Political Liberalism*. New York: Columbia University Press.
———. 1999a. *A Theory of Justice*. Cambridge, Massachusetts: Belknap Press.
———. 1999b. Kantian Constructivism in Moral Theory. In *John Rawls: Collected Papers*, ed. S. Freeman. Cambridge, Massachusetts: Harvard University Press.
———. 1999c. Themes from Kant's Moral Philosophy. *John Rawls: Collected Papers*, ed. S. Freeman. Cambridge, Massachusetts: Harvard University Press.
———. 2007. *Lectures on the History of Political Philosophy*. Ed. Samuel Freeman. Cambridge, Massachusetts: Belknap Press.
Sandel, M. 1982. *Liberalism and the Limits of Justice*. Cambridge: Cambridge University Press.
Scanlon, T. 2003. Rawls on Justification. In *The Cambridge Companion to Rawls*, ed. S. Freeman. Cambridge: Cambridge University Press.

Public Reason 1 (2): 31-44

Keeping Truth Safe From Democracy

Christopher Jay
University College London

Abstract. The ambition of 'justifying democracy' has more and less theoretical aspects, involving more or less emphasis on our pre-philosophical commitments. The prospects for justifying democracy without recourse to pre-philosophical commitments are not good, as we see if we are properly critical of David Estlund's admirable recent contribution to the democracy literature. What does this mean for political philosophy? We must think hard about the role of justificatory projects with the ambition of doing without pre-philosophical commitments. Such projects are not without a role, but it is a constrained role.

Key words: justifying democracy, Estlund, pre-philosophical commitments, political theory.

There are deep issues about the extent to which our pre-philosophical intuitions and commitments ought to constrain our philosophical theorising about politics. Sometimes, as in the work of liberals such as Rawls, this manifests itself as a question about how members of a pluralistic society, in which individuals and groups have different attitudes and values, and adopt various 'comprehensive doctrines' governing their thinking about moral and political questions some of which will be pre-philosophical might settle questions about what the basic structure of society ought to be. Rawls's own strategy amounts to focussing on ways in which agreement about the most important questions of justice might be reached from such disparate viewpoints. In the course of reaching agreement, intuitions are brought into 'reflective equilibrium' with more considered, theoretical principles, and therefore play a central role in determining the eventual shape of a liberal theory of justice.

Another way in which pre-philosophical ideas about politics can enter into political philosophy is less explicit. When political philosophers turn to questions of state legitimacy and democracy, they might conceive their task as one of establishing the *theoretical grounds* upon which to rest the sorts of political arrangements we pre-philosophically approve. For such a philosopher, the task is not to rethink *whether,* say, democracy is good, but to say *why* it is good. This way of approaching political philosophy serves an important purpose: one of the things we all ought to want our intellectuals to do is to make clearer to us the real contours of our existing commitments, the better to understand the values which in some sense make us what we are. (This might, in line with the political liberal's conception, amount to understanding a range of values, or a disjunctive set of values.)

But there are other ambitions which the political philosopher might reasonably have. We might, for example, want to know not just what is good about democracy, but whether it is really as good as our pre-philosophical ideas would have it. (This might be out of purely philosophical interest, or it might be because, for example, we are sensitive to the historical delicacy of democratic ideals such as ours.) And we had better be careful not to allow these distinct – though both important – ambitions for political philosophy to

become confused, on pain of diluting the rigour proper to both ambitions. We had better not miss what is important *to us* about our pre-philosophical ideas when pursuing the first aim, by confusing what an idea *is* with what we find appealing *about* it. And we had better not let our pre-philosophical ideas play too great a role in our theorising when what we are putting on trial are those pre-philosophical ideas themselves. Our pre-philosophical ideas can, when we are pursuing the second aim, be called as witnesses but must not be admitted to the jury.

How, though, are we to say when our pre-philosophical ideas have ended up playing too much of a role? This is bound to be a delicate question. I propose a case study, in the form of a discussion of an excellent recent contribution to the philosophical debate about democracy. I shall conclude with some general suggestions about what this case study tells us about when our pre-theoretical ideas might, and when they might not, be important for political philosophy.

I. DEMOCRACY

Why might we think that democratic institutions and practices are the best ones to adopt or approve of? There are two sorts of answers one might give, and which have been offered in the literature on the foundations of democracy: it could be that democracy does uniquely well at respecting, exemplifying or promoting some particular values; or it could be that democracy promises to be an optimally efficient problem solving mechanism. There are, of course, many variants of each broad strategy and much to say about them. But I want to discuss a specific attempt, in David Estlund's (2008) book *Democratic Authority: A Philosophical Framework*, to describe what we might think of as a 'mixed' justification,[1] and proceed to say some things about the prospects for a theoretical defence of democracy more broadly, given the lessons of what I will suggest are Estlund's insights and his eventual failure.

Estlund's view is that democracy is the most efficient problem solving mechanism *which meets a criterion based on respecting a particular value*. Specifically, Estlund thinks that we should expect democracy to do well at coming up with political decisions which avoid what he calls "primary bads" (war, famine, economic or political collapse, epidemic and genocide), but that a condition on any justified political authority is that it is acceptable to those who are subject to that authority.[2] There might be even better ways to arrive at good political decisions, by "epistocracy" – rule by the wise – most obviously, but they fail the acceptability condition and are therefore ruled out.

1] All parenthetic page references in the main body of the text are to this work.

2] Estlund's discussion of 'authority' is extensive and interesting, but I shall not address it here and will assume that it will be clear from context what the notion of authority is as I am employing it.

II. SAVING POLITICAL TRUTH, BUT AVOIDING EPISTOCRACY

One of the most appealing aspects of Estlund's view is that he resists the idea that there are no political truths (or none in a 'robust sense') and that there is no such thing as political expertise. Several writers have proposed that since there are no robust facts that determine political truth or expertise we are forced to approve of democratic mechanisms on the grounds that they are uniquely well-suited to deciding questions by surveying myriad 'truths' or conceptions of the truth.[3] By associating political truth with the prevention of primary bads, Estlund has pointed out a fairly straightforward way of seeing that scepticism about truth in politics cannot be right: political truths need not be an exotic variety of truth, nor even necessarily as problematic a variety of truth as moral truth; rather, political truths will just be the truths about what must be done, or how things need to be, if we are to achieve whichever political aims we have.[4] Deciding about which aims we *should* have will involve difficult value judgements (and we might be sceptical about the conditions for problem solving adequacy being met by democracy here), but even if scepticism about value facts is reasonable, there is no reason to think that all political facts are value facts. The point is that, unless we are willing to accept some form of global scepticism about facts or truth, we have no reason to be worried about the existence of at least some *political* facts or truth, and therefore some political expertise.[5]

3] See e.g. Botwinick 1990 or Hatab 1995. Richard Rorty (e.g. 1991 [1988]) seems to come close to the view I have in mind here.

4] Estland also, therefore, avoids the position on truth in politics associated with political applications of classical pragmatism as found in Misak 2000, Talisse 2005 and – in a slightly different way – Habermas (e.g. 1996 [1992]), which seek to rethink what the notion of political truth (indeed, truth itself) amounts to instead of seeking to do without any talk of it.

5] But note that the writers I referred to in n3 (and to a lesser or more problematic extent n4), above, are attracted to some form of global scepticism about 'robust' truth in the most everyday sense. This does not mean, however, that they all reject objectivity or embrace relativism (see, e.g., Misak's 2008 defence of objectivity with respect to moral deliberation and personal experience). At this point it is as well to be careful about what is at issue here. I do not mean to suggest that Estlund, in adopting talk of primary bads for what I am arguing are the purposes of saving the notion of political truth, is committed to any disagreement with Rawls (his avowed inspiration) and in particular the Rawlsian denial of the role of truth in favour of reasonableness. Rawls, as became clear for example in the course of his debate with Habermas (cf. Habermas 1999a and 1999b), thought of the 'political not metaphysical' conception of liberalism thus: "Political liberalism does not use the concept of moral truth applied to its own political (always moral) judgements. Here it says that political judgements are reasonable or unreasonable; and it lays out political ideals, principles, and standards as criteria of the reasonable." (2005 [1995]: 394). This explicit rejection of truth as the operative notion for liberalism throws welcome light on comments in *A Theory of Justice* such as "granting that God's will should be followed and the truth recognized does not as yet define a principle of adjudication" (Rawls 1999: 191). And I take it that Estlund would not disagree: we might paraphrase Rawls on his behalf and say that "granting that some measure will decrease the likelihood or severity of some primary bads does not as yet define a principle of adjudication". Rawls would not appear (from his 'granting' the premise) to want to deny that there might be a truth of the matter about the need to follow God's will (though even with the benefit of the later clarification it is still unclear what the force of this 'need' is supposed to be), just as Estlund does not want to deny that there are truths about primary bads. And just

Avoiding Estlund's primary bads is one way of being concerned with quite straightforward political truths, but there are others. Whether raising interest rates will effect inflation is, perhaps, a political fact (since it is a fact about politically relevant institutions and conditions), as is whether the offending rate for some particular crime(s) is falling or rising, and whether some particular measure is likely to effect this. And as Estlund accepts, if there are political truths it makes sense to think that there is – or at least *could be* – political expertise. Certainly the possibility of political expertise is not *entailed* by the existence of political truths or facts, for there could be epistemic obstacles to grasping the truths that there are. But in the absence of an argument or some evidence to motivate the idea that such epistemic obstacles stand in our way, or at least that what obstacles do stand in our way are *particularly* problematic, it is reasonable to expect that the non-exotic political truths at least will be susceptible to the same sorts of investigation as any other sorts of facts about which we readily accept that there might be experts. Estlund is right to say that these are relevant considerations when thinking about justifications for political decision making arrangements: if there is political expertise, it seems quite reasonable to include the means of exploiting that expertise amongst the desiderata for our political institutions and practices.

Where Estlund errs is in his rejection of what he acknowledges to be the natural conclusion from these considerations. His acceptance condition on political authority rules out a government of the wise on the grounds that (i) even genuine experts might not be *recognised as* experts by those who will be subject to their authority (so they will quite reasonably withhold their acceptance of the experts' authority believing the experts, mistakenly, to be unqualified), and that (ii) it would be reasonable for those whose acceptance is required to withhold it on the grounds that bias (even unforeseen bias) might creep into the decisions made by any select (and relatively homogenous) group of decision makers.[6] I think both these arguments are specious, and that the acceptability requirement itself is under-motivated.

I will discuss the requirement itself below, but first I shall say something (briefly) about each of Estlund's two specific worries. Both seem to be just as worrying for the democrat as for the "epistocrat". As for the first, the likely failure of a significant proportion of the population to recognise expertise just looks like grist to the anti-democrat's mill: do we really want to endorse the practice of open elections to positions of power and influence if recognising expertise is going to be a problem? Perhaps the problem is still worse for the democrat: with respect to complicated issues (such as economic policy and analysis, or international relations) recognising competency might be just as tricky as recognising expertise. And perhaps for the most important positions of responsibility expertise is what is required for competency – to be a competent foreign affairs advisor,

as for Rawls this truth would apparently not suffice to "define a principle of adjudication", for Estlund the requirement to avoid primary bads is presumably a tenet of reasonableness, not a brute matter of fact.

6] These considerations are also discussed in Estlund 1993 and 2003 respectively.

you had better be an expert on some aspect of foreign affairs. The democrat's position would not be helped by arguing that the role of elected representatives is to represent the people's opinions or interests, not to uncover objective political truths, since knowing what the real opinions or interests of a constituency are and what implications they have for policy decisions is a form of political expertise (there is an objective fact of the matter about what the opinions or interests of a constituency are – this is, for example, why polling is a difficult science), so we have just as much reason to be sceptical that people will be good at recognising it as to be sceptical of their ability to reliably track any other skill. As for the second of Estlund's worries, the familiar tendency of democracy to favour populist measures and to foster climates of debate in which only a quite restricted range of opinions and argumentative strategies are admissible strongly suggests that if there is a worry about some particular decision making procedures fostering decisions biased towards some particular range of interests, then that should be a worry for the democrat just as much as for anyone else. Estlund might think it is especially problematic that the rule of the wise would risk legislating in favour of minority interests (those of the wise). But I do not see how this could be the right distinction: it cannot just be that minority interests are specially unfit for legislating in favour of because they are the in the minority – that is just to assert the democratic thought which these considerations are supposed to be grounding, so a circle beckons.

So much for the idea *that* epistocracy fares worse on acceptability grounds. Estlund's discussion of *why* his acceptability condition is supposed to make more trouble for epistocracy than for democracy is also rather unconvincing. His argument is that, whilst democratic arrangements are not the default preferable ones, any authority requires justification (hence the application of the acceptability condition), and the democratic ideal is one in which nobody has authority over any anybody else, thus placing it in at least *prima facie* pole position if all appeals to authority of some over others fail (36-38). I do not see the force of Estlund's claim that democracy does not involve authority of some over others: Mill's worries about the "tyranny of the majority" seem reasonable, and seem reasonable precisely because the majority have authority over minorities in a democracy.[7] (Notice that however many minority rights are protected in the sense familiar in the context of, e.g., ethnic minority protection, the very fact that democracy hands decision making priority to majority opinions or preferences is sufficient to ensure that majorities have

7] See Mill 1859. Incidentally, Mill did not think (what is frequently attributed to him) that the educated should have plural votes so as to outvote the uneducated majority. Rather (see Mill 1861, §8), his explicit preference is for the educated to have just as many plural votes as is necessary to even up the voting profile in light of the fact that the uneducated might, by sheer weight of numbers, systematically outvote the better educated. Mill's concern is clearly with the problematic ignoring of specifically educated voters, but his concern for minority interests per se was sufficient to limit his plural voting principle so as not to create a new voting minority to be systematically outvoted, namely the uneducated. So it is clear, I think, that the real force of Mill's worry about the tyranny of the majority applies wherever majorities and minorities are at stake.

authority over minorities, where these groups are defined in terms of their opinions or preferences, regardless of the constraints there might be on the exercise of that authority.)

III. THE QUALIFIED ACCEPTABILITY CONDITION

My brief discussion of Estlund's *application* of his condition was intended to show how, taking the idea of acceptability as an intuitive test for legitimacy (not, that is, as a test with strictly defined rules), Estlund has succeeded in identifying some of the more pressing issues to do with political authority, but has not drawn the most *prima facie* satisfying conclusions from them. This, I suggest, means that unless his *particular* brand of acceptability condition can be shown to have some independent grounding we have reason to accept his premises (to do with truth, expertise, the recognition of experts and decision making bias) but reject his conclusion (that epistocracy fails, in favour of democracy).

So I shall now turn to Estlund's acceptability condition itself. On Estlund's picture of his hybrid of procedural (value-exemplifying) and epistemic (problem solving) criteria, his procedural criterion (the qualified acceptability condition) acts as a sort of brake on his epistemic criterion (the avoidance of primary bads):

> [T]he bindingness and legitimacy of the decisions are not owed to the correctness of the decisions, but to the kind of procedure that produced them. Still a central feature of the procedure in virtue of which it has this significance is its epistemic value … Democratically produced laws are legitimate and authoritative because they are produced by a procedure with a tendency to make correct decisions … [D]emocracy is better than random and is epistemically the best among those that are generally acceptable in the way that political legitimacy requires. (8)

What precisely is the "qualified acceptability condition"? Estlund's idea of the "necessary condition on the legitimate exercise of political power" is "that it be justifiable in terms acceptable to all qualified points of view (where 'qualified' will be filled in by 'reasonable' or some such thing)" (41). Such a condition speaks to the "Expert/Boss Fallacy":

> [F]rom the fact, even granting this it is a fact, that you know better than the rest of us what should be done, it certainly does not follow in any obvious way that you may rule, or that anyone has a duty to obey you … To the person who knows better, the other might hope to say, "you might be right, but who made you boss?" (40)

The Expert/Boss Fallacy seems to be problematic because of the "Rawlsian thought" that "it would be a kind of intolerance to think that any doctrines could form a part of political justification even if some citizens conscientiously held reasonable moral, religious, or philosophical views that conflicted with them" (43-44).

But it is not quite clear what would drive us, on theoretical grounds, to accept the acceptability condition. The Expert/Boss Fallacy is only a fallacy if it is *false* that knowing better *makes* you boss, and it is not obviously false. At least, there are contexts in which it seems clear that expertise implies authority: when caught up in a traffic accident in the street, for example, the mere fact that one person knows about first aid *is* usually suffi-

cient to just make them boss, at least with respect to what the rest of us should be doing to help.[8] The falsity of the thesis that knowing makes you boss – the falsity of the "epistocracy thesis", that is – is just what the acceptability condition is supposed to entail, so its falsity should *not* be what's motivating the acceptability condition itself.

Connectedly, it is not at all obvious why the 'kind of intolerance' involved in the Rawlsian thought is problematic. It would be problematic if that intolerance failed some condition, such as the acceptability condition. But without assuming the acceptability condition it is not clear which condition it is supposed to fail; and of course we cannot just assume the acceptability condition since, if the acceptability condition is supposed to be motivated by the need to avoid the intolerance, a circle beckons. Even if reasonable alternative conceptions could be specified and sorted from unreasonable ones, it remains to be seen what is supposed to ground the inference from a view's being reasonable to it being a transgression of some condition not to tolerate it. There seems to be a *normative* gap in the argument thus far. Apart from anything else it seems rather implausible that, with decisions needing to be made and deliberation having to conclude some time, the fact that some others reasonably disagree with some decision is sufficient to render implementation of that decision problematically intolerant[9] – and that is just the sort of situation at issue, plausibly, in the debate about the virtues of democracy. So it is unclear what is supposed to move us to agree that it is, in general, *problematic* that we are intolerant of differing reasonable opinions, so long as intolerance is distinguished from persecution. (It is obvious that whether or not we tolerate dissenting opinions by according them meaningful input into the decision making process, it is a further thing to punish or disadvantage someone *for holding* a dissenting opinion. I have no doubt that this *further* thing is wrong.)

Estlund defends his acceptability condition against two sorts of objection, which he dubs the 'over-*ex*clusion' and 'over-*in*clusion' objections. His responses to both are clever, but they betray the extent to which Estlund's official theoretical defence of democracy in fact rests unduly upon his *pre-theoretical* concern to avoid epistocracy.

The over-*ex*clusion objection, according to Estlund, charges that the qualified acceptability requirement rules out too much, specifically that it rules out certain (i) *possible*

8] This is, I think, a quite reasonable suggestion in its own right. But as it happens Estlund appears to agree that (in an example similar in all relevant respects) the authority of the expert is justified with or without the acceptance of that authority (or 'consent', as he calls it in his discussion): see Estlund 2005 §IV. In fact, I think Estlund's views about the acceptability condition are in serious tension with his excellent discussion of "null" non-consent in the first half of that essay, though he appeals in the second half to a controversial idea of the non-transitivity of authority conditions which might rescue the acceptability condition (though he does not say this in the essay). But I will confine my discussion here to his Democratic Authority view.

9] Of course Estlund, presumably invoking Rawls, would almost certainly retort that what is at stake with the Rawlsian thought is not the problematic intolerance involved in doing something that other people reasonably object to, but with failing to take those objections seriously. But quite what taking them seriously amounts to is not clear, unless we assume the democratic conception of political participation, which conception we are, of course, trying to ground by means of this very thought. Again, the circle beckons.

objections that people subject to some decision *might* have, and/or certain (ii) *actual* objections that people subject to some decision *do* have. But, Estlund maintains, neither (i) nor (ii) are problematic for him: (i) ruling in all *possible* objections is absurd, not because of the sceptical result that there would be no authoritative political principles, but because such a sceptical result should not be entailed by the merely logical fact that any principle *can* be negated (that is, objected to); and (ii) ruling in (all) *actual* objections is entirely consistent with his qualified acceptability requirement – qualified acceptability is merely a *necessary* condition for authority, not sufficient, so it is left open whether there are further conditions, such as actual acceptance (44-49). I suspect that Estlund would eventually recoil from allowing actual acceptance as another necessary condition on legitimacy, but in any case I think his response to the over-exclusion objection is uncharitable. I suggested above that there are cases in which we do not object to epistocracy or intolerance. Why is that a worry? Because it is not clear what the salient differences between those cases and the political case are – is it that in the counterexample cases no "qualified" (reasonable?) person would object? Estlund explicitly avoids specifying what "qualification" amounts to, but we would know more if we knew which (sorts of) *possible* and which (sorts of) *actual* objections were qualified. That, I take it, is the force of the over-exclusion objection: it enquires as to which objections are qualified and which are not, by suggesting that all are qualified and challenging Estlund to say which are not. Estlund assumes that the objection is a positive thesis claiming that *all* possible and/or actual objections should be counted and that it is a problem that his requirement doesn't count them; but, if it *is* such a positive thesis, then it is better seen as a provocation to Estlund to say more about exactly which cases the objection gets wrong – *that* will be its real force. Estlund's points in response don't say anything more (at least anything which addresses this issue), so we still don't know what separates the counterexamples from the political case. The powerful over-exclusion objection, then, is that Estlund seems to exclude too many grounds for withholding consent since excluding *any* grounds for withholding consent without explaining why some are excluded and not others is illegitimate. (We should always worry when a distinction looks arbitrary.)

The over-*in*clusion objection is supposed to go like this: since justifications based on true premises and sound reasoning are successful on just those logical grounds, they establish legitimacy regardless of who might object. Estlund asks us to consider such an argument (50):

(P1) Christianity is a truth of the utmost importance;

(P2) Truths of the utmost importance ought to be taught in state schools;

(C) Therefore, Christianity ought to be taught in state schools,

Suppose that (P1) and (P2) are true. That, the over-inclusion objection says, is sufficient for the truth of (C), *regardless of who agrees or disagrees*. But that, Estlund argues, is not good enough to show that the acceptability requirement rules out too much:

> I grant that [C] follows from [P1] and [P2], regardless of who might disagree. Does this establish the over-inclusion objection? … The answer is plainly 'no.' … The dispute [in question] is not about whether valid arguments from true premises establish their conclusions … [but rather] the truth about legitimacy. Premise [P2] makes a claim about legitimacy that is not obvious, and is denied by the qualified acceptability requirement. (51)

The issue is not, Estlund thinks, between those (his opponents) who value truth and those (Estlund) who supposedly cleave to some other criterion:

> [I]f you love the truth, then you want to know what account of legitimate coercion is true. One possibility, the view taken by the qualified acceptability requirement, is that the true view says that political justifications are specious if they appeal to doctrines that are not acceptable to all qualified, even if mistaken, points of view. (51-2)

But:

> Nothing I have said shows that the qualified acceptability requirement, rather than the exclusive view, is true. My aim is only to point out that it, too, would be a truth. (52)

None of these observations seem particularly dangerous to those sympathetic to the over-inclusion objection. It is clearly true that the point at issue between Estlund and his opponent here is to do with the truth of any premise which asserts the legitimacy of non-acceptable conditions or arrangements. But noting that that is what is at issue gives his opponent no reason at all to recoil from simply maintaining that such premises are *true*. Estlund has said that he considers them false, but as the debate stands it resides in stalemate. It is plausible that there might well be terms on which the issue could be decided. Indeed, Estlund probably thinks of his "Rawlsian thought" as precisely the sort of way in which to decide the issue, modulo my charge of under-motivation. So the over-inclusion objection is not properly countered until we have been given some (good) *reason* to agree with Estlund that the controversial premise is false.

IV. KEEPING POLITICAL PHILOSOPHY SAFE FROM INTUITIONS, AND INTUITIONS SAFE FOR POLITICAL PHILOSOPHY

So, what do I take my comments about Estlund's view to show? I think that the acceptability condition and its supposedly motivating Rawlsian thought about intolerance raise a question about why we should accept them, specifically that it is not at all clear where to draw the line between cases in which expertise does entail authority and those in which is does not. If there is no line to be drawn, then the Expert/Boss Fallacy is not a fallacy at all. But knowing where to draw the line would require committing to some robust theory about the relation between expertise and authority. Estlund offers no such account (his account is not robust, since it does not tell us how to decide where to draw

the line), and proceeds on the assumption that it will be obvious how to sort cases. His responses to the over-*ex*clusion objection – which is most charitably interpreted as pressing on just this issue – and the over-*in*clusion objection betray the extent to which he seems content to rely on his democratic intuitions to ground his arguments. But of course it was the legitimacy of those democratic intuitions which was at stake in the first place, for the defender of epistocracy need not deny that we have democratic *intuitions* but might charge that they are misguided.

As Estlund argues (ch. 10), procedural approaches to justifying democracy tend to end up appealing to epistemic (problem solving) criteria in the end,[10] so we should not think that our search for a philosophical grounding for our democratic intuitions should just turn to the procedural strand – there is every reason to think that similar problems (problems with scepticism about democracy on epistemic grounds) will recur. And it seems even more problematic to turn to a purer form of epistemic justification – it was the threat to democracy from purely epistemic criteria that forced Estlund to introduce his acceptability condition in the first place. So it seems that reflecting on the issues discussed above raises important questions and doubts about the project of justifying democracy at all (at least if the taxonomy I have assumed – epistemic justification, procedural justification, or Estlund-style mixed justification – is exhaustive).

The justificatory question, about *whether* and *why* we might think that democratic institutions and practices are the best ones to adopt or approve of, occupies an interesting place in the various debates about democracy that go on both within and outside the academy. It is not immediately obvious that those of us debating democracy in the comfortable surroundings of broadly democratic states have any reason to worry about whether and why democracy is a good thing, rather then just how best to make sense of it and make it optimally virtuous. But at the same time we frequently come face to face with issues which make the justificatory question – theoretical as it is – pressing and important.

The thought that those of us who enjoy the obvious (though not unqualified) benefits of living in broadly democratic states have little reason to worry about the theoretical question of justifying democracy follows from reflecting upon the role of theorizing about politics. On one quite plausible view, our theorizing should serve the purpose of explicating the assumptions and intuitions which – at some sort of deep or fundamental level – we share. On this view, there is both a contingency and necessity to our assumptions and intuitions: their contingency is recognised by prescinding from seeking to justify or ground them in such a way as to rule out all possible alternatives (which would be trying

10] Estlund argues that models of deliberative democracy such as those offered by Jurgen Harbermas or Joshua Cohen rely on notions of an 'ideal' deliberative situation that determines an ideal solution against which actual deliberation is to be judged, which solution is as independent of actual deliberation as any objective problem solving notion offered by the epistemic conception. And I think that, on pain of implausibility, any conception of democracy (so any supposedly grounded in procedural justifications) must recognise the importance of epistemic criteria when deciding issues to do with who gets to participate, i.e. to be allowed membership of the demos (see Dahl 1979).

to dig below bedrock); and their necessity by taking them to be basic enough to build upon as axioms for our political thinking (taking them to *be* bedrock, for us). This view, then, suggests that the failure of the justificatory projects discussed above is of limited importance or interest for our serious thinking about real political issues and attitudes: Estlund's retreat to pre-theoretical presumption in favour of democracy is not a problem for the justificatory project, rather it is just what sensible justification should be built on.

This view is not straightforwardly wrong since even those of us who, donning our philosopher's hats, explicitly reject the justificatory arguments for democracy (even if we were to explicitly reject the priority of democracy itself) still *think like democrats*.[11] That is, our coming to be well-adjusted members of democratic communities just is, in an important sense, our coming to adopt or accept broadly pro-attitudes towards broadly democratic institutions and practices, and towards ideals (inclusion, respect, participation…) which, if not essentially democratic, are essential to democracy and, on the broadest construal of democracy, part of the democratic ideal. Even those of us who reject the ideal tend, if we are well-adjusted members of communities for whom democracy *is* an ideal, to manifest (probably unthinkingly in all but the rarest cases of explicit theorising) at least most of the basic moral attitudes typical of fully paid-up ideal democrats. So there must be more at stake for us, as members of democratic communities, when thinking about democracy than just the truth or falsity of some propositions or the success or failure of a justificatory project: at stake *for us* is whether we are really ready or able to follow the impeccable logic of our theorising so stringently as to change not just what we think, but also *what we are*.[12] Since it is unlikely that we really *are* ready to do the latter in all seriousness, there is *something* to the thought that the *useful* role of theorising about democracy is constrained by our basic attitudes and the principles and ideas we 'deeply accept'.

But that view is, I think, only partially right. The view of theorising I just described as not straightforwardly wrong is not straightforwardly right, either. Our theorising about our own democratic institutions and practises might well be constrained by the background noise of our deeply held assumptions and strongly felt attitudes, but there are cases where the very nature of a political problem or issue strongly recommends (perhaps even *demands*) that we address it theoretically in just the way pursued above if we are to reach a settled view at all. So, for example, we might not seriously propose to overthrow our own basically democratic existing system of government on the basis of our discovery (if it is a discovery) that there is no philosophically coherent theoretical basis for any presumption in favour of democracy which doesn't appeal to our pre-theoretical commitments, but we might very well think that such a discovery would be a good reason for prescinding from

11] Note that I am not suggesting that we all think as democrats, even if we avow a rejection of the democratic ideal; rather, I am suggesting that were we to reject the democratic idea we would nonetheless literally think like democrats – we are not all democrats, but even those of us who are not (but who are 'well-adjusted' to our socio-political surroundings) systematically think in similar ways to those who are democrats.

12] To the extent that what we are is determined by the attitudes we adopt to ideals and principles.

demanding *more* democracy (even if it is not a good reason to demand *less*). This is not mysterious: whenever our deeply held beliefs or most strongly felt attitudes run short of offering a conclusion (short, that is, of offering up a settled view with the elements of contingency and necessity described above), some further investigation is appropriate and, as in the example just given, it is likely that questions going beyond the broadly democratic *status quo* will go some way beyond the constraints of our socialised background noise of pro-attitudes and beliefs.[13] Similarly, if the question is whether some state which is not democratic would be better for being democratic, whether we should 'export' democracy or encourage it, it seems obvious that our democracy-friendly background pro-attitudes and beliefs will be, at least, subject to a more disinterested theoretical perspective – it is not our own situation at issue, after all, and so the issue of whether, at the terminus of enquiry, we are ready to change not only what we think but *what we are* does not arise, or at least does not arise in the same vivid way: what *we* are to be is not at issue, does not constrain our theorising in the same way. And since our own existing practices are not at issue, it is likely (again, as in the previous example) that our most basic pro-attitudes and beliefs will underdetermine strong settled views about the political plight of others (particularly of others whose non-democratic situation is not obviously worse than our own in humanitarian terms). It is these sorts of cases, then, that show the role that purely theoretical enquiry plays in our thinking.

So, to bring the strands together, it seems that the failure of Estlund's justificatory project is instructive because it highlights a tension in democratic theory: there are political truths and political expertise, and these facts are not irrelevant to the theoretical justification of any political system. But democracy does not come off well with respect to those facts – it is likely that some sort of epistocracy would do better. So if there were to be a theoretical presumption in favour of democracy, it would (as Estlund sees) have to be that epistocracy fails some condition which democracy does not. Estlund, I argued, fails to motivate such a condition, so epistocracy remains a real theoretical alternative to democracy. My particular worries about Estlund's strategy – which, I think, is the most thoughtful strategy currently on offer, so the best test-case for the hope of justifying democracy – converged on the worry that at crucial moments Estlund's 'justification' rests on his pre-theoretical presumption in favour of democracy. The subsequent discussion of the relation of theorising to our pre-theoretical commitments was in order to place this worry about Estlund (and, since he is our 'test case', about the theoretical justification

13] Note that I am not suggesting that theoretical enquiry is only appropriate where no pre-theoretical attitudes or beliefs would suffice for reaching a settled view. What I am suggesting is that there are some attitudes and beliefs which, in virtue of their being the very attitudes and beliefs the acceptance of which constitutes our growing into well-adjusted members of democratic communities, are sufficiently 'deep' or important for us in terms of our being what we are that they trump – for better or worse – disinterested theoretical enquiry when it comes to thinking about certain things. As it goes, I do not think that the set of such deeply accepted attitudes and beliefs is very large, so I don't think that what I am suggesting here represents any sort of general threat to the value of theory or of theoretical enquiry in general.

of democracy more broadly) in the context of our political thinking more generally. If what I said about that is right, then I think something like the following picture emerges. The bleak prospects for a theoretical justification of democracy which does not rest upon pre-theoretical democratic commitments will not be particularly significant for our most basic settled views about our own broadly democratic societies – at least, it is unlikely that anyone will be persuaded by my criticisms of the justificatory project that the basic structure of their own broadly democratic society needs to change. But the failure of the justificatory project *will* be significant with respect to other issues, namely (among others) whether our institutions or practices should change not in their basic structure but in their details so as to make them more democratic, and whether non-democratic others would be better off as democrats.[14] The failure of the justificatory project will be significant here because theoretical reasoning is required to justify, to ourselves and each other, our conclusions in light of the fact that the questions at issue seem genuinely open, unconstrained by the worrying prospect of having to change what we are by acting on some particular conclusions. Keeping truth safe from democracy – respecting political truth to the detriment of giving priority to democracy – remains, if my comments on Estlund are right, theoretically desirable; *that* fact, if it is a fact, might not be expected to bear upon our attitudes towards our own broadly democratic basic structure, but should be significant for thinking about *how much* democracy we want, or whether it would be best for others.[15]

c.jay@ucl.ac.uk

14] I concentrate upon debates about whether we ought to have more democracy because it is rare these days to come across serious proposals for less democracy within the broadly democratic structure. That is, it tends to be that – given the assumption that the broadly democratic structure of political life is the right sort of structure – arguments are advanced to the effect that (or, often, from the assumption that) we ought to make particular institutions (the House of Lords in the UK or the European Union, for example) more democratic. It is a delicate question whether debates about whether particular institutions ought to be less democratic – whilst still granting the desirability of the broadly democratic structure – would, if they were to be found, be subject to the same degree of underdetermination by our 'deep' values as I have suggested debates about more institutional democracy are. It is also a delicate question precisely to what extent debates about whether democracy is best for others tend to be infused with considerations of whether the sorts of people we are – democrats, or those who think like democrats – need to approve of (and/or work to achieve) democracy for others, in order to be properly 'authentic' democrats ourselves. This will, of course, have implications for what I have been claiming; but deciding that issue involves interesting work for elsewhere.

15] An ancestor of this paper was discussed at a UCL seminar in spring 2008. Thanks to all those who attended, and particularly to George Hull, Rory Madden, Mark McBride, Poly Pantelides, Arthur Schipper and Jose Zalebardo for comments, questions and/or criticisms. Thanks also to Dave Holly and Craig French for discussion of various issues overlapping with the topic(s) of this paper. At the University of Brighton CAPPE conference 'What's the Big Deal about Democracy?' another version received an airing and was the subject of comments from Alison Assiter, Hanno Birken-Bertsch and Jurgen de Wispelaere, for which I am grateful.

REFERENCES

Botwinick, Aryeh. 1990. *Skepticism and Political Participation*. Philadelphia: Temple University Press.

Dahl, Robert A. 1979. Procedural Demcracy. In *Philosophy, Politics and Society* 5th series, ed. Peter Laslett & James S. Fishkin. Blackwell Publishers.

Estlund, David. 1993. Making Truth Safe for Democracy. In *The Idea of Democracy*, ed. David Copp, Jean Hampton, and John Roemer. Cambridge: Cambridge University Press.

———. 2003. Why Not Epistocracy? In *Desire, Identity and Existence: Essays in Honor of T. M. Penner,* ed. Naomi Reshotko. Academic Printing and Publishing.

———. 2005. Political Authority and the Tyranny of Non-Consent. *Philosophical Issues, 15: Normativity*.

———. 2008. *Democratic Authority: A Philosophical Framework*. Princeton: Princeton University Press.

Habermas, Jürgen. 1996 [1992]. *Between Facts and Norms*. Trans. William Rehg. Cambridge: Polity Press.

———. 1999a [1996]. Reconciliation through the Public Use of Reason, reprinted in Habermas, 1999c.

———. 1999b [1996]. '"Reasonable" verses "True," or the Morality of Worldviews', reprinted in Habermas 1999c.

———. 1999c [1996]. *The Inclusion of the Other: Studies in Political Theory*. Trans. and ed. Ciaran Cronin and Pablo De Greiff.

Hatab, Lawrence J. 1995. *A Nietzschean Defence of Democracy: An Experiment in Postmodern Politics*. Chicago: Open Court.

Mill, John Stuart. [1859]. *On Liberty*. Reprinted in Mill, 1993.

———. [1861]. *Considerations on Representative Government*. Reprinted in Mill, 1993.

———. 1993. *Utilitarianism, On Liberty, Considerations on Representative Government*. London: Everyman.

Misak, Cheryl. 2000. *Truth, Politics, Morality: Pragmatism and Deliberation*. London: Routledge.

———. 2008. Experience, Narrative, and Ethical Deliberation. *Ethics* 118 (July)

Rawls, John. 1999. *A Theory of Justice* Revised Edition. Oxford: Oxford University Press.

———. 2005 [1995]. Reply to Habermas. Reprinted in *Political Liberalism* Expanded Edition, New York: Columbia University Press.

Rorty, Richard. 1991 [1988]. The Priority of Democracy to Philosophy. Reprinted in *Objectivity, Relativism and Truth: Philosophical Papers*. Cambridge: Cambridge University Press.

Talisse, Robert B. 2005. *Democracy after Liberalism: Pragmatism and Deliberative Politics*. Oxford: Routledge.

Public Reason 1 (2): 45-72

Of Human Bonding:
An Essay on the Natural History of Agency

Mariam Thalos & Chrisoula Andreou[1]
University of Utah

Abstract. We seek to illuminate the prevalence of cooperation among biologically unrelated individuals via an analysis of agency that recognizes the possibility of bonding and challenges the common view that agency is invariably an individual-level affair. Via bonding, a single individual's behavior patterns or programs are altered so as to facilitate the formation, on at least some occasions, of a larger entity to whom is attributable the coordination of the component entities. Some of these larger entities will qualify as agents in their own right, even when the comprising entities also qualify as agents. In light of the many possibilities that humans actually enjoy for entering into numerous bonding schemes, and the extent to which they avail themselves of these possibilities, there is no basis for the assumption that cooperative behavior must ultimately emerge as either altruistic or self-interested; it can instead be the product of collective agency.

Key words: agency, altruism, bonding, collective action, cooperation, emotional attachment, identification, individualism, self-interest, team reasoning.

"Life is a long trip in a cheap car. In a dark country. Without a good map."[2]

The topic of what sustains cooperation among biologically unrelated individuals in evolutionary history knows no disciplinary boundaries. How do we explain the evolution and stability of such instances of cooperation? This question was heralded appreciatively in the present scientifically-minded era by E. O. Wilson in his controversial classic *Sociobiology*, although Thomas Hobbes was acquainted with a certain, unbiologized version of it much earlier. And the question remains with us still. Ranged among those who find it within the ambit of their concerns one can find psychologists, biologists, anthropologists, linguists, political theorists, students of computation, game theorists, and of course philosophers. The question has generated a wealth of cross-disciplinary conversation that promises to impact public policy palpably. For if it is determined that the correct answer to the question is that individuals cooperate (or comply with an unfavorable condition) only if doing so serves them severally according to a favorable pay-back schedule, then public policies will be framed accordingly: the success of public policies will be estimated according to the schedule of incentives they promise to complying individuals. This is already the foundation of many economists' proposals and theories – to the chagrin of those with more optimism about the better angels of human nature. But of course if the more pessimistic position on the question of how cooperation is sustained is in fact bet-

1] We are indebted to Lije Millgram and Nick White for thoughtful comments and conversations.

2] Opening words to Frederick Schick's *Making Choices* (1997), written with the aim of improving Bayesian decision theory.

ter supported by the evidence, the economists in question are not simply jaundiced but rather proceeding on the best possible grounds.

We shall argue that scientific theories contending that cooperation among strangers rests ultimately on a foundation of self-service, or service of one's lineage, are ill-founded, and indeed ill-supported by experimental evidence. For these theories are founded on an aprioristic restriction of the search space to mechanisms of launching behavior at an individual level. This restriction of the search space functions as a highly problematic assumption to the effect that motivation is an individual-level affair. We shall be challenging this dogma. To be sure, we are not the first to challenge the individualistic dogma, nor are we contending that all scientific theories on this topic are guilty of the individualism we shall be indicting. What we are offering here is a systematic criticism, across a range of literatures, and advancing also a corrective that helps to cast nonindividualistic proposals in a new light.

We shall be arguing that prior to the question of the evolution of cooperation is a parallel but different question that must be answered. This is the question of the evolution of agency itself: how did agency, as such, emerge on the evolutionary landscape, and in such units as we actually find it on the ground today?

Now, an agent is a unit – indeed a *unity* – that takes (or at least launches) action; an agent is related to its deeds as author and not merely or necessarily as proximate cause. To be sure, it is an empirical question whether a given thing, indeed anything at all, qualifies as an agent in this sense. For about a century now the wisdom among those with naturalistic inclinations has been that the idea of an agent is a construct that has no role to play in naturalistic explanations. Relatedly, much of psychological theory in the last century consistently treats the "self" as simply a collection of so-called *self-beliefs* or *self-attributions*, a collection of self-relevant beliefs, rather than as something that could legitimately be treated as an entity in its own right.[3]

There is of course a clear advantage to those conducting primarily psychological or behavioral science, as well as to philosophers following their progress, of conceptualizing agency in terms of performance criteria: this is that it is unproblematic to examine and document an organism's cognitive performance on tasks, enumerate and diversify the tasks examined, and then proclaim when once enough of these are co-present in an organism, that the organism qualifies as an agent (or at least as having a precise quantity of intelligence, that seems itself to be a place-holder for agency).[4] And it is quite possible that a critical mass of such capabilities will be sufficient to guarantee the presence of agency

3] See for example the essays arrayed in Wegner and Vallacher (1980) as well as the pieces in Duval, Silvia and Lalwani (2001), in which "self," as such, is never distinguished from "self-concept," "self-awareness" or "self-standards". Contrast this with new and important research on self-regulation prominently led by Charles Carver and Michael Scheier (2001), in which the "self" is construed as a sum total of self-regulation processes; and cf. contributions by Demetriou and Kazi (2001).

4] The strategy is illustrated admirably in Byrne (1995), but Byrne makes no contentions vis-a-vis agency as such.

as well. But it is very questionable to suppose that such a critical mass, despite being sufficient to guarantee the presence of agency in prototypes of the species, is necessary for possessing agency, or sufficient in every case. More problematic still is the supposition that a critical mass in ability to perform such tasks itself *amounts to* or *constitutes* possession of agency. The most obvious reason for denying this latter supposition is that execution of the tasks in question might be organized in a distributed or decentralized fashion, and that a disunified aggregation of performances, however expertly carried off, does not add up to an agent – which, by definition, is a unity.

This state of affairs is perhaps largely responsible for the sustained flowering of transcendentalism. Dissatisfaction with the naturalists' treatment of the agent has contributed to a growing sense that science does not – and cannot – give an adequate or complete treatment of agency, because the reality of agency, as such, *transcends* the methods of science. Transcendentalists defend the existence of a special sphere or realm of which science cannot treat: the first-personal, transcendental world of the Self and Others, which has at least since Kant been considered the private preserve of Philosophy with a capital P. Advocates of this transcendentalism usually insist that we *not* call the study of the transcendental sphere a science. These intellectuals do not suffer from science envy; they are card-carrying Philosophers. Science, according to transcendental philosophy, can take us only so far in the intellectual journey. And the suggestion is that perhaps Reason, with a capital R, or simply Intellect, can take us beyond the frontiers of science.

In our view, social scientific research on agency has tended to define its search space too restrictively, and there is no reason to deny that agency has a role to play in naturalistic explanations. Now of course there are important details to work out – most notably, handling the question of what qualifies an entity as an agent. And of course there is no shortage of answers to this question on the transcendental side. The trouble for our purposes is that a preponderance of published opinions on this subject, on both sides of the transcendental divide, simply assumes that the boundaries of agents coincide with their skin, or their fur or what-have-you.[5] On what foundation does this assumption repose? On nothing but philosophical (decidedly even political, and in particular neoliberal) dogma, as will become clear. As we will show, the answer to what agency consists in must be responsive to empirical findings about the behavior of contemporary humans, as well as to evolutionary considerations, and these empirical findings contradict the dogma of whose truth we are constantly being assured. The failure of this dogma, as we shall argue, is reason to be suspicious of current wisdom on the cooperation question. We will show that numerous incidents of cooperation among the unrelated are inexplicable by a calculus of self-service, and instead are better explained by a calculus whose subject is some "we." We will provide a taxonomy of "we"-s that distinguishes among its targets of analysis on the basis of how they are forged, which will provide a natural taxonomy for types of cooperation.

5] Indeed one of us has proposed an alternative conceptualization of agency that deliberately shuns this assumption: Thalos (2007; 1999), cf. also Thalos (2008).

Before we begin it will be well to handle in advance one obvious initial reaction to our proposal. Notice first that we couch our contention in terms of the entity level at which identification of motivation is appropriate. A critic might reply that motivation is in no way the issue here, that what is at stake is the level at which interests are served, or simply at which advantage accrues, and that any talk of motivation has been purely incidental or metaphorical, simply a rhetorical device for marking interests. Economists, evolutionary biologists, psychologists and ecologists – among others – have settled on what may at first glance seem a sterile or bloodless way of handling explanation of the individual human behaviors they aspire to explain. They view individual human behavior as best modeled on a fiction – namely, that the behavior is undertaken as a means to solving a certain decision problem, through maximizing a return on some investment. They do not view the individuals that manifest the behavior as *themselves* carrying out the calculation that solves the decision problem. Nor do they view these individuals as themselves understanding their situation in terms of a decision problem that (first) calls for a cost-benefit calculation, and (subsequently) leads to a motivation in favor of the option that wins the day. Therein lies the fiction: there is no real-time decision processing in the proper sense of the term; indeed, there may be no cogitation of any kind. Nonetheless they view the behavior as *best explained* on a model that weighs certain costs against certain benefits.

We grant the propriety of this form of explanation; in fact, it is one of our own contentions that the issue of agency is prior to and in many ways independent of psychological mechanisms that today underlie motivation. And in drawing attention to motivation we too are drawing attention to the level at which advantage accrues, contending that agency issues are intertwined with advantage and that evolutionary pressures can be brought to bear upon units of agency themselves.

Why then draw attention to the topic of motivation? Why not retreat to a *generalist* position on the subject of explaining behavior?[6] The generalist position seeks to explain behavior by directing attention purely to advantages conferred by the behavior on those entities engaging in it, without making any attempt to identify the relevant psychological machinery for controlling that behavior. The issue would then coincide exactly to the issue of levels of biological selection.

What the generalist proposal lacks in detail it makes up for as follows. The generalist proposal is simply to identify behavior that can outperform a range of competing behaviors (in evolutionary terms, during a certain period in evolutionary history), and subsequently to claim that this advantage makes the behavior inevitable even for those of us who come so much later. Advocates of this generalist position argue that a less generalist account, that deals in psychological details, might obscure this fact. It might suggest that details matter, when they don't. For if the behavior weren't achieved through the particular ways it was achieved, it would have been achieved some other way. And this is the generalist's point. We could subsequently cast our contentions in terms of the level at which

6] This is Robert Batterman's term (1998, 76-102) to refer to the work of Robert Axelrod (1984) and Brian Skyrms (1994, 305-320; 1996).

interests accrue, and leave out discussion of motivation entirely. That would serve certain of our immediate ends. But it is unsatisfactory in the long run to leave out issues of motivation, because in the end the agent is an entity defined at least in part by how it navigates in relation to motivations, its own and others'. And so to say that agency emerged in response to an evolutionary pressure is to say that unified and motivated entities so emerged. And with this statement comes the obligation to develop an account of how agency manifests itself in developmental time, and how it is transmitted down a lineage.

According to our account, motivation is a matter of being drawn to a goal or object – a feature abstracted enough from any physical realization of it to be realizable by a group or collection of dispersed biological individuals. A great deal hangs on what we count as agents (especially as increasingly much of what human life depends upon is determined by what happens to common-pool resources). And so any contentions about what may so qualify should be supported with argumentation and not merely subjected to fiat. This is the imperative to which this paper is responding. It will proceed in a way that treats agents as natural kinds, asking but not purporting any decisive answers as to how units of action can emerge in evolutionary time.

I. THE QUINTESSENTIAL SOCIAL SPECIES

"Sex," as E. O. Wilson (1975) remarks, "is an antisocial force in evolution." It constrains large-scale organization and division of labor – not because it interferes with labor or the individual variations among organisms that makes division of labor efficient, as quite the contrary is true – but because it stands in the way of unproblematic division (however inequitable) of the fruit of such labor. With the rare exceptions of monozygotic (identical) siblings, no two organisms in a species that reproduces sexually are genetically identical. Conflict – if only in strictly biological, reproductive terms – is therefore inevitable. To the extent that more for your interests means less for mine, collaboration between the two of us cannot be without its strains. Whereas, by contrast, where there is no genetic gap between the two of us, your interests and mine will coincide exactly, and there is no sense in which more for your interests means less for mine. Our cooperation in such an instance can proceed without hesitations.

Thus one way of overcoming obstacles to large-scale cooperation is to close the genetic gap, as has been done in the social insects of the order Hymenoptera via the device of *haplodiploidy*.[7] With sterile castes and suppression of reproduction among females, the gap

7] This is the mechanism by which males (developing from unfertilized eggs) have only one copy of each chromosome (haploid), while females are wrought the usual way and with two copies of each chromosome. This mechanism has important consequences: a queen's daughters from the same mating (called *supersisters*) are highly related to each other, and a female is more related to her sisters (on average 75%) than she is to her own daughters (on average 50%). Thus haplodiploidy opens the way for a worker caste, devoted to helping their mother. Sterility is a superior strategy when it is more expedient (less costly) to help a mother beget a sister (or many sisters) than to have a daughter of one's own. See Wilson and Hölldobler (1990) and (2009).

between individual interest and collective interest among the social insects is appreciably closed, making the family or colony the (veritably one and only) unit on which the forces of natural selection act. As R. A. Fisher remarked, "The insect society more resembles a single animal body than a human society…The reproduction of the whole organism is confined to specialized reproductive tissue, whilst the remainder of the body…tak[es] no part in reproduction" (1958, 200). In pronominal terms, social life among the social insects is a matter of "we" (at home) and "they" (when one colony encounters another). "I" and "Thou" have no genuine place.

By contrast the gap of genetic relatedness between organisms in a modern human community is as large as it can be among sexually reproducing species. Still, human societies enjoy (if anything) a run-away division-of-labor that leads to continuously escalating organizational structures entirely unprecedented on the planet. Human beings today live in large settlements, many of them phenomenally large, covering vast territories, and comprising numerous genetic lineages of unrelated individuals. According to Wilson's ground-breaking 1975 treatise, human societies comprise one of four pinnacles of social evolution, and cannot be explained entirely by mechanisms that support the welfare of close kin – mechanisms that ably explain the other three pinnacles of social evolution: colonial invertebrates, eusocial insects and nonhuman mammalian societies. As early as 100,000 years ago, humans lived in hunter-gatherer family units tied by cooperative bonds at a tribal scale, having no more in common than language and distant common ancestors. Even the simplest contemporary tribal societies link family units of a few tens to create societies of a few hundred to a few thousand, held together by common sentiments of membership, "expressed and reinforced by informal institutions of sharing, gift giving, ritual, and participation in dangerous collective exploits" (Richerson & Boyd 1999, 254). These tribal ties – very possibly constituting a necessary developmental stage along the way to large settlement living – are unprecedented in evolutionary history.[8] How is this form of social organization – which we will refer to as *network society* – to be explained?

Full-scale settlement living has many advantages: easy resource defense and reduced vulnerability to predation are perhaps the least controversial. So it might seem reasonable to propose that social living amongst humans, and the precursory network societies, evolved precisely because of or for these advantages. But there are problems with this proposal: for there also are disadvantages to large settlement living, with enormous susceptibility to disease being among the most important.[9] How do the advantages of settlement living measure up against the disadvantages?

8] Richerson and Boyd write: "We know of no close analog of tribes in other species" (2001, 211); they also write that "larger, more complex societies are generally able to dominate smaller, simpler tribal societies, and a ragged but persistent trajectory of social evolution toward ever more complex social systems continues to the present" (1999, 254).

9] Diamond (1999) powerfully documents the impact of contagious disease upon human societies in an important class of cases.

One suggestive strategy is to enumerate more and more advantages (and disadvantages too) to large settlement living – more than the obvious ones – and argue that together these constitute an enormously favorable balance over hunter-gatherer ways of life in nuclear family units. This strategy might work, if we could be assured that the evolution or development of the newly enumerated advantages does not require that large settlement living, and solutions to the problems and risks associated therewith, precede them. For example, it might be reasonable to suppose that settlement living makes possible the production of a food surplus, and so allows a large group to stockpile against the risk of famine. But does food surplus production require that large settlement living already be long and firmly established? If so, then we may have part of the story concerning the *expansion* of settlement living, but no part of the story concerning its *establishment*.

Eventually, the problem with the strategy of simply enumerating the advantages of large settlement living is that as we multiply purported advantages, ultimately we will run headlong into the problems of unrelatedness: what are the advantages that accrue to unrelated individuals through (for instance) the division of labor? For it might well be that the division of labor is more advantageous to some individuals than to others, for whom it might be a considerable burden. Unlike reduced susceptibility to predation, not everyone benefits to the same degree from the division of labor, since the surplus from it is rarely divided evenly. Indeed, some may not benefit at all – they might actually suffer in relation to how they might have fared with less social organization.

The point here is that the move to cooperative network living among the unrelated cannot be conceived entirely as a solution to a problem of pure coordination; for unlike solutions to problems of pure coordination, community living is not always win-win, or at least not obviously so. The mixed blessings of the likes of the division of labor, for example, are just as liable to destabilize the growth of settlements as they are liable to foster it. Thomas Hobbes knew something of this reality when he wrote that, in the state of nature, "there is no place for industry, because the fruit thereof is uncertain, and consequently, no culture of the earth" (1994 [1668], 76). To be sure, while the blessings wrought by a food surplus are blessings indeed, if I am not so positioned within the network society as to be confident I will be enjoying them, why should *I* count them as advantages of settlement living?

And so we arrive, finally, at a fundamental record-keeping question – the question to which we will devote the efforts of this essay: to whom is settlement living (or anything else) supposed to be advantageous? Must we view advantages as accruing, always and everywhere, to individuals, considered individually, or can our record books contain entries for groups, considered as groups over and above the individuals that comprise them? If the latter, then it might turn out that settlement living is a mixed blessing for both types of entries. As we will see, this is an important issue, as it serves to illuminate the question of just how an advantage functions in the logic of a purported explanation of its contemporary prevalence. We shall make room for answering the question vis-à-vis associations of

the unrelated in much the same terms as Wilson answers the question vis-à-vis associations of the very tightly related.

Whatever must be said about the balance of advantages to disadvantages, complex social organization in large settlements (as well as those things that come with it: domestication of plants and animals, large-scale food production, technological innovation of all varieties, writing systems and systems of communication generally) has grown up semi-independently more than once in the human lineage, and developed along a variety of different lines, supported by a variety of different social arrangements and cognitive structures.[10] So there must be a basis for it. And we'd like to understand what that basis is in as general terms as possible.

It is important to emphasize two facets to this theoretical problem. The first is largely historical and ecological: how did complex social organization arise among humans in the first instance? Unique features of species are typically the result of the niche the species occupies – its biogeographical location. What specific ecological *problems* did complex social organization solve for hominids? What are the ecological precursors and prompts? Did the organization in question arise in stages? If so, what were these stages, and what the nature of progression through them? Must all societies that reach this or a similar complexity in their social organization take this same route to it? (In other words, is this destination reachable only by one developmental route, or are there alternates?) This facet of the problem concerns the dynamics of the development of social organization, in relation to ecological conditions that favor or impede it.

The second facet of the theoretical problem is less historical and more ontological: it is concerned with the nature of the mechanisms and the entities that sustain and maintain the novel social organizations, and the boundaries that these things impose upon potential developments. What is it about the entities themselves, and the ways they function and develop, that sustains the structure in question, and constrains its further development and evolution? Are the mechanisms entirely biological? Or perhaps they are psychological in character? Maybe they are purely social, or a combination of all three. Are they present in their entirety in (some or all) of the individuals whose populations enjoy the organizational complexity? Or are they features of populations as populations?

Evolutionary game theorists are concerned with the broadly historical/dynamical questions. Theirs is the generalist approach. Other research, some of it conducted in the area of evolutionary psychology, is concerned with sustaining mechanisms.[11] Ultimately, the research in these two areas must be synthesized. For if, on the one hand, we go without a treatment of the ontological problem, it would appear that we have no more than

10] Nisbett (2003) argues that East Asian cultures differ from western cultures in profound ways that are anchored in cognitive strategies.

11] Barkow, Cosmides, and Tooby (1995); Buss (1994). Other sources are: Buss and Malamuth (1996), and Crawford and Krebs (1998).

mythologies or just-so storytelling.[12] If, on the other hand, we go without treatment of the historical/ecological problem, we do not have clear confidence that the proposed mechanisms explain what needs explaining, and whether they are indeed the best candidates for explaining it.

Though it might be too ambitious to address the two facets of the problem simultaneously, each facet can be addressed with some sensitivity to the other. Our purpose in this essay is to enlarge the toolkit for modeling the evolution of cooperation on the ontological side of the question, by taking steps to develop of a taxonomy of bonding schemes among organisms. Via bonding, a single individual's behavior patterns or programs are altered so as to facilitate the formation, on at least some occasions, of a larger entity to whom is attributable the coordination of the component entities. Some of these larger entities will qualify as agents in their own right, even when the comprising entities also qualify as agents.

II. THE MULTIGENERATIONAL DIMENSION: IN THE BEGINNING

Why do organisms bother investing in reproduction? The answer, of course, is that those who fail to do so die without descendants, and only those who manage to do so maintain a lineage. Among plants, those that manage to invest resources into reproductive functions win; and similarly in the animal kingdom. But among members of the animal kingdom, some young require more than just the opportunity to live. They cannot simply 'take it from there'; instead, they require some care to actualize their reproductive potential. Parents who fail to provide such young with the necessary care, leave no descendants into third and fourth generations. So how could species with such young have evolved in the first place? Nature, as we will now discuss, has provided these organisms numerous ways of bonding with their young.

Mammals give birth to live young. And so mammal mothers can meet their offspring face to face, without much risk of mistaking the relationship. This provides an opportunity that nature can take advantage of to reward those mammals that provide an optimum of postpartum care for their young. Mammal young need to be nursed for a period of time; mammal mothers who fail to nurse their young run the risk of leaving no progeny behind. Those who provide normally do so by way of having bonded with those young. The more bonded, the more they provide. Homo sapiens, like all other mammals, some birds and some reptiles, are a bonding species. Bonding is a solution to an evolutionary

12] One profound philosophical worry about purely dynamical approaches to the ecological problem is that they might be largely irrelevant. The concern is that the origin of any phenomenon consists in a trajectory of unique historical events, and so it is subject to empirical inquiries – anthropological, archaeological and sociological. An investigation of purely evolutionary-dynamical issues may reveal that certain historical sequences are ruled out, but it's not likely to reveal a unique sequence as the actual one. And so anthropological, archaeological and other empirical inquiries would still be required to discriminate the most likely among eligible historical trajectories.

problem – indeed the most fundamental one. Not surprisingly then, bonding is a universal of mammalian life, and must come on the scene long before you and I are considering whether to live together as cooperating strangers.

This truth has yet to be integrated into explanations concerning how cooperative living amongst unrelated individuals comes to prevail, and how it is sustained and multiplied. Yet the answers to the questions of (1) why (and how) organisms provision their young, and (2) why (and how) unrelated or distantly related organisms provision one another, are fundamentally related. They are related as species of the same genus, as our proposal will explain.

It is now commonplace to view the structure of the ecological problem faced by would-be cooperators as a Prisoners' Dilemma (PD). Early explicit expressions of this idea are due to John Maynard Smith (1982), Robert Axelrod (1984), and William Hamilton (Axelrod and Hamilton 1988), who introduced the apparatus of game theory into biology, and evolutionary biology in particular. But germs of the idea go back as far at as Thomas Hobbes. This game-theoretical conceptualization is combined with the assumption that cooperative living amongst unrelated individuals cannot be explained in essentially the same way that parental investment in mammals is explained – namely, via an appeal to bonding.[13] This assumption is, we will argue, mistaken – bonding is a key to cooperative living even amongst unrelated individuals. This suggestion is bound to seem implausible if one supposes that all bonding takes the form of emotional attachment; but emotional attachment is just one of a variety of species of bonding, as we will now explain.

III. BONDING IS A GENUS

In this section we shall be offering a taxonomy of bonding. Before we offer our three primary taxa, we will do well to mention some forms of association that might, under the right circumstances, qualify as precursors, if not also as bonding taxa in their own right.

Coral polyps live in colonies. Each polyp benefits, perhaps in only a very small way, from the proximity of the others. Perhaps the sheer number of cohabiting polyps helps to create a more stable habitat, moderating to some extent environmental variables like temperature. Assuming this is all that the colony provides the individual, and assuming that coral polyps have to affix themselves somewhere and together is just modestly better than apart, this seems a rather flimsy basis for thinking there is something that deserves calling coordination of behavior here. Still, because polyps possess little or no locomotion, it would seem that their "choice" of a home is everything in the world that they are entitled to calling behavior, and so it might qualify after all. If in the end we view this example and

13] Recently Skyrms (2003) has added the stag hunt game as a conceptualization of the ecological problem. His model is also founded on the assumption that cooperative living amongst unrelated individuals cannot be explained via an appeal to bonding.

others like it as continuous with the others, then we shall perhaps wish to add another taxon to our taxonomy. But we shall leave that issue open for now.

Symbiotic relations are everywhere. In mutualism, a dyad is formed whose members are generally of different species, each benefit from a close ecological association. The association might be negotiated by a coordination of behaviors, or simply as a result of happenstance that each finds itself in the right place at the right time. If it is through the former route – through coordination of behavior in one of the three ways we will be discussing – it will make sense to refer to the dyad as a bonded entity.

Consider now the following three species of bonding:

Imprinting

Two individuals can become a type of natural unit via the process of imprinting. Goslings, for example, imprint on the first moving creature they see (usually their mother), and then constantly follow this creature around. To take a somewhat different example, when a mother ewe gives birth, she imprints on the smell of her baby's wool while licking off the amniotic fluid covering the newborn. Within five minutes "the door to maternal tolerance slams shut" and the mother rejects any baby that does not smell exactly like the baby (or babies) she has imprinted on (Hrdy 1999, 158).

Given the right conditions, imprinting effectively attaches mother and offspring, which helps explain why it is sometimes selected for. The point is simply that the behavior pattern is triggered by a cue and is from that point forward comparatively rigid. This fact explains the imprinting "errors" that have been documented: The first moving creature a gosling encounters may not be its own mother, but a curious researcher, and bonding might occur anyway. And a mother ewe can be 'tricked' into adopting a lamb that is not her own if the lamb is smeared with her fresh amniotic fluid.

Emotional Attachment

A different, more familiar form of bonding for us humans is bonding via emotional attachment. Bonding via emotional attachment is more gradual and less mechanistic than imprinting. It typically requires, and is reinforced by, extended contact or 'face time.' In evolutionary terms, bonding via emotional attachment has its advantages. Like imprinting, emotional attachment can serve to bond kin to one another. But unlike imprinting, it admits of degree, and thereby allows for commitment to vary with emotional investment. (This can result in a variation of attachment strengths, depending upon expected payoffs of the attachment.) For example, while a mother ape will not care for a newborn that cannot cling to her, she will care for an infant to which she has become attached even if it becomes too weak to cling. (This makes sense, as the evolutionary value of a sick 6-month-old is considerably higher than that of a newborn, provided its chances of recovery are good.)

But again there is room (and in fact more latitude here) for copying errors, because even in the absence of blood kinship and potential for reciprocity, contact can breed emotional attachment. And so sacrifices for genetically unrelated individuals can be rampant in environments in which contact between genetically unrelated individuals is rampant. (In one clear instance of this in our own society, emotional attachment to adopted infants explains investment and even sacrifice between the genetically unrelated.)

Identifying

Another familiar form of bonding for human beings is bonding via identification. Our cognitive capacities allow us to recognize bonding and to think in terms of "we"; relatedly, they allow us to theorize or postulate "we"-s and to act accordingly. In some cases, the recognition of similarities suffices to get individuals postulating a "we" and thinking in terms of their (collective) good (Kramer and Brewer 1986); this, in turn, may be enough to sustain their existence as a "we." When it comes to identification, what is logically essential is not the development of certain emotions or *feeling*, but rather the development of a certain *conception* of things.

Note that identification serves to unify not only groups, but also individuals. (Indeed, development of the first personal "I" is perhaps the first appearance of bonding of this kind.) At least typically, when one takes into account how one's current actions will affect one's future self, one does this out of one's identification with one's future self. Although one sometimes reasons from the point of view of oneself-now, one usually reasons from the point of view of oneself-as-a-temporally-extended-being or from the point of view of some "we" of which one is part. Such reasoning involves identification, but need not involve emotion; in particular, it need not involve a feeling of sympathy. For example, it is my identification with my future self, rather than a feeling of sympathy for my future self, that gets me to the grocery store on Saturday mornings regardless of whether I am feeling hungry.[14]

Notice that it is possible to postulate a "we" and identify with others based on a faulty conception of one's relationship with those others. One may, for example, think of oneself as part of a team, even though no such team exists because the other supposed team members do not think of themselves as part of a team. But even when one correctly thinks of oneself as part of a team, this may seem like a mistake in evolutionary terms (some will insist) if those with whom one identifies are, by and large, not kin. Such behavioral developments should not (or so one might argue) be carried forward into future generations. For identification supports sacrifice, and sacrifice for non-kin is not generally fitness-enhanc-

14] Nagel (1970) argues that reason requires that agents take the interests of their future selves and of others into account. Given the notion of identification, one might interpret this as suggesting that agents must (on pain of irrationality) identify with their future selves and with others. This position is bolder and more questionable than our own. All we are suggesting is that agents (at least human agents) *can* identify with their future selves and with others.

ing. And even if identification can be fitness-enhancing in our own era, where advantages can accumulate to identifiers in numerous ways, how can it have evolved in the first place?

IV. WHY ARE WE HYPER-SOCIAL?

In this section we shall draw together the main elements of our proposal, contrasting it with our competitors' while drawing the relevant parallels between bonded groups and individual organisms. The question we are trying to answer is of course: Why – and how – are we humans so hyper-social? The major competitor to our proposal is this one: We are hyper-social through reading each other's minds (this answers the "how" question; Baron-cohen 1995, Tomasello 1999), and this is so that (now turning to the "why" question) each of us can keep from falling behind in strategic advantage in a competitive world characterized by shifting alliances (Byrne and Whiten 1985). And so it's all for individual advantage. Less starkly put, because each of us has the capacity for understanding the minds of others, through being able to put oneself cognitively and emotionally/motivationally in another's place, we have been able to learn from one another, play with one another, share information and ultimately live together in (apparently) cooperative communities.

This proposal rests on an apparent consensus, forged among philosophers and scientists of cognition, around the Representational Mind, which is the hero of the so-called "Machiavellian intelligence hypothesis" (Byrne and Whiten 1985). The ecological challenge, on this proposal, is to predict the behavior of others, and this requires reading their minds so as to represent it to oneself in a kind of practical syllogism performed third-personally. Once one's mind becomes readable, one has to raise the stakes by learning to conceal one's mind as well as to penetrate through such concealments as others find advantageous to perpetrate, in an escalating arms race of deceptions and unmaskings.

One version of this proposal (due to Hrdy 1999, 2009) declines to some extent the "Machiavellian intelligence" aspect of this hypothesis: No, it is not for warfare's sake (either for concealing one's intentions or for unmasking others' intentions) that we understand each other's minds; it is rather so that each individual among us (adult and immature alike) can find and/or please *allomothers* – adult care providers and protectors, which are indispensable because maternal commitment among humans is much more conditional than among other primates; but, once again, it's all for the sake of individual profit.

One challenge to the entire project of Machiavellian intelligence seeks to provide an alternative model of cognition to explain all the phenomena of hypersociality, without reference to hyperactive Representational Minds.[15] For instance, Strum, Forster and

15] Behaviorists and cognitivists alike, pervasively both within the discipline of psychology and outside it, have used the term "agent" to refer to any entity with the representational properties of mind, without giving due consideration to whether there is an open question as to the connection between goal-orientation on the one hand, and representational states on the other. This is evident in the title of an essay – indeed, an entire volume of essays – by the prominent philosopher of biology Kim Sterelny, *The Evolution*

Hutchins (1997) argue for a distributed cognition model of social processing, seeking to "undermine the very narrow individualistic language of tactics and strategies and the disembodied view of cognition that have been the basis of our approach to the primate mind in the new cognitivist era" (1997, 73). But even here there is no forceful challenge of the notion that the gold standard for explaining behavior in an individual is *individual* advantage. To be sure, there is a suggestion that a larger "cognitive system" may be in evidence. But there is no suggestion that natural selection acts on characteristics of *it,* favoring some characteristics and not others.

Our proposal, by contrast, is this: We are hyper-social through having managed to bond in multiple ways (answering the "how" question), which has led to the adoption of multiple collective goals which result in an overall reduction in conflict of interest; and all this is so that (turning now to the "why" question) the units that have bonded together can outperform other competing units. This, in broad outline, is the analysis we find E. O. Wilson and Bert Hölldobler (both 1990 and 2009) making vis-à-vis cooperation in the order *hymenoptera.* And correspondingly, we make no commitments as to the physical realization of goal orientation among the bonded – whether via a Representational Mind or something else – just as Wilson and Hölldobler makes no such commitment as to the physical realization of superorganismic bonding among ants, although they are keenly aware of the existence of chemical signaling among them (and motor programs among honey bees). We believe that the issue of what is in an individual's mind in a moment of hyper-sociality, just as the issue of the chemical signals or dances exchanged in ant and bee communication, is in fact a red herring. What is of significance is the *communication-structuring* motivation. And this may be nowise evident in human minds as such, just as the analogous articles may be nowise evident in the chemical signals exchanged by ants.[16]

Wilson and Hölldobler (2009, 6-7) describe the construction, in natural history, of organism and "superorganism" as perfectly in parallel. Just as there is no question as to what is in the "mind" of a single cell when it contributes to the behavior of the organism it helps to compose, so also: "Nothing in the brain of a worker ant represents a blueprint of the social order…Instead, colony life is the product of self-organization…The assembly instructions the organisms follow are the developmental algorithms" (2009, 7). And these are simply the result of natural selection operating in the usual way, but on the superorganism as a whole.

---

of Agency (2003), in which the only entities that the title could possibly be naming are organisms purportedly in possession of folk-psychological states of believing and desiring; nowhere in the book does Sterelny acknowledge a need to give further account of agency, but takes it simply for granted that "agency" and "thought-out behavior" are co-referential.

16] In accordance with the observations in the previous footnote, philosophers who cannot separate the notion of agency from that of a compendium of representational states – as cognitivists nowadays cannot – will be inclined to respond here that, accordingly, we must not be going on about anything recognizable as agency. We reply that, to the extent that the concept of *motivation* is itself inseparable from that of agency, we are indeed going on about something recognizable as agency. And we remain neutral here as to the relationship of motivation and mental states.

Like Wilson and Hölldobler, we view bonding as the result of natural selection operating in the usual way, upon characteristics of the higher-level bonded entities themselves, rather than upon the characteristics of the composing individuals therein bonded. If this bonding proposal is correct, it might explain a great deal – some of which cannot be explained at all by competitor proposals. In the following section, we shall, by drawing on empirical work, elaborate on those phenomena which our bonding proposal is better-placed to explain; and in subsequent sections we shall further contrast our proposal with other bonding-like proposals.

V. IDENTIFICATORY BONDING BREEDS COOPERATION IN THE LABORATORY

Identificatory bonders are capable of assimilating themselves into larger "we"-units. And, insofar as they are capable of reasoning and acting, they are capable of reasoning and acting as part of larger "we"-units. They can be found asking themselves "What should we do?" and "What would be best for us?" (These are all too familiar questions within the context of family units.) When such questions arise, it is often obvious that *we* will do best if each of us cooperates with the others, and so our deliberation is focused *not* on *whether* to cooperate with one another, but on which of the options available to us as cooperators to aim at. And importantly: when an individual acts as part of a "we", that individual is not acting simply on reasons had as an individual – that individual is acting out of reasons had as a part of a larger entity.

Bonds can, of course, vary in strength (as they do for example in molecules). Where bonds are very strong, bonders will reliably make significant sacrifices in order to do what is best for the group. Where bonds are weak, bonders may be willing to make only small sacrifices for the group's sake.

Consider the following set of experimental results described in Dawes *et al.* (1997). In one experiment (run by A. J. C. van de Kragt, R. M. Dawes, and J. M. Orbell with S. R. Braver and L. A. Wilson), groups of subjects were put in PDs. Some of the groups were given ten minutes of discussion time before each participant had to decide what to do. Other groups were not given the opportunity to communicate. Each participant then gave her confidential decision to the experimenters before leaving. While the average cooperation rate in the groups with discussion time was about 80%, the average cooperation rate in groups with no discussion time was about 40%. These results led to the hypothesis that group solidarity might be very important for reliable cooperativeness. Two follow-up experiments sought to determine whether conscience or the opportunity to make commitments, rather than solidarity, could account for the increased cooperativeness among groups with discussion time.

In one of the follow-up experiments (run by Orbell, van de Kragt, and Dawes), all participants had some discussion time; but in some cases, the discussion was with the group that would determine the participant's payoff and benefit from any contribution

she made; while in other cases, the discussion was with participants whose decisions would not affect her payoff or benefit from any contributions she made. The idea was that

> if discussion triggers conscience, and our contributing subjects are acting to satisfy its demands, then discussion should enhance contribution to strangers. If, however, discussion elicits caring about group members, then it should enhance contributions only to people in the group with whom one interacts. (384-5)

The researchers found that, "contrary to the clear conscience hypothesis," discussion "does *not* enhance contribution when beneficiaries are strangers" (387).

In another follow-up experiment (also run by Orbell, van de Kragt, and Dawes), groups were monitored and any verbal commitments made were tracked. Researchers found that, except where every member in the group made a verbal commitment, there was no relationship between promising and actual choice. It was concluded that, while promises are effective in "universal promising groups," it is "solidarity – not commitments per se – that leads to the higher level of cooperation in [these] groups" (389).

Note that, as Dawes, van de Kragt, and Orbell point out, although group interaction can effectively generate group solidarity, interaction between group members is not essential for the generation of group solidarity. It has, for example, been demonstrated that individuals tend to exhibit solidarity with respect to individuals whom they do not know (and whom they have not had any contact with) but whom they think of as sharing their fate (Kramer and Brewer 1986).

The suggestion that rational players in a true PD might rationally cooperate is, to certain prominent game-theoretic minds, little short of apostasy. It can only proceed, as Ken Binmore writes, from "the wrong analysis of the wrong game" (1994, 114). For, according to Binmore, it follows from the very meaning of the notion of 'payoff' taken together with the fundamental game-theoretic conception of rationality that a rational player in a PD must choose to defect (and hence if a rational player does not do so, it was not a PD in the first place). The idea that players in a PD might give some pride of place to *sums* of payoffs, adding in payoffs not reflective of their individual concerns – and suggestive instead of some "we" – would suggest to Binmore and others that players in the dilemmas construed by Dawes and Orbell do not perceive themselves to be playing a PD. (To a first approximation, they are construing the situation as an iterated PD.)

To explain why players in real-life experiments (both in and out of the laboratory) frequently choose to contribute to a collective good when faced with dilemmas whose material payoffs accord with the PD,[17] many strategies have been devised that involve the transformation of material payoffs into utilities whose structure deviate from the PD sufficiently as to make the observed behavior consistent with the game-theoretic solutions.[18]

17] A meta-analysis in Sally (1995, 62) reveals that summing across all 130 PD experiments carried out between 1958 and 1992, the proportion of subjects choosing cooperation over noncooperation is 47.4 percent.

18] A useful cross section of this literature is surveyed in Hargreaves-Heap & Varoufakis (2004, ch. 7).

Specific implementations of this include ascribing to players a dislike of or aversion to inequality, or ascribing to players a liking for reciprocation (or retaliation); either way, the result is a boost in subjective attractiveness, to the relevant players, of cooperative outcomes.[19]

What critics of cooperative solutions to the PD as such are unwilling to acknowledge is this reality: when we put subjects in a PD-like situation, by fixing their payoff schedule accordingly, we aren't specifying that they must construe their dilemmas as dilemmas for individuals or dilemmas for groups (nondeliberating, to be sure, and possibly also scattered in space and time). Indeed we cannot do so: it is an experimental manipulation that is simply out of reach. But if participants elect the latter construal, neither decision theory as such, nor any of its axioms individually, can censure them for it. So, if subjects are indeed irrational to construe themselves as groups in such circumstances as we place them, it is not decision theory that can convict them of it. Is such a construal irrational? Perhaps, at least in certain cases, natural selection might be in a position to impeach them for a misstep in construal – or at any rate to punish them. But its verdicts are not to the effect of "irrational!" For "Team Think" is not – as such – irrational, even if it might be in some other way practically inadvisable under certain circumstances.

Michael Bacharach has been developing "Team Think" analysis of cooperation in PD cases.[20] On his analysis, a player in a PD can either construe the dilemma as an individualist or else favor a collective construal of it. But on Bacharach's analysis, this is not a further choice – so that the collective computation is no intermediate step towards an ultimately individualistic construal. On Bacharach's account, "team reasoning" is fundamentally opposed to individual reasoning, so that if you are conducting deliberations in the one idiom, you cannot rationally be conducting it also in the other. Any given concrete dilemma is "spontaneously" framed *either* as an individual choice or as a collective one – never both at the same time. But frames can vary. Someone, on his view, reasons as a team member if she chooses the act (if this is unique) that is her component of the profile that (as she has worked out) is best for the objectives of some group (1999, 32). This reasoning, as Bacharach maintains, "is a basic decision-making proclivity of mankind; … it is fundamental to the workings of organizations of diverse forms; … it is a concomitant of group identification;…and … it completes the theory that group identification is the basis proximal mechanism for successful human group activity" (2006, 121).

19] For specifics, see for example, Fehr and Schmidt (1999), Bolton & Ockenfels (2000), Geanakoplos et al. (1989), and Rabin (1993).

20] Unfortunately, Bacharach is not clear on the point of whether team reasoners in situations where their payoffs can be characterized by a PD matrix are really in true PDs. His untimely death in 2002 was a tragic loss to the discipline, as he was still in the process of writing a book that pulled together the themes on foundations of decision theory that he had been developing for decades. Fortunately he had completed enough of this book that we can get a picture with broad strokes (but which is ambiguous on the point Binmore stresses). Thanks to Robert Sugden and Natalie Gold for taking on this important work: Bacharach (2006).

For our purposes, it doesn't matter whether team analysis of the PD schedule of payoffs is conceptualized as fundamentally inconsistent with individualistic analysis. What matters for our purposes is whether the process of team analysis brings into existence a new agency that vies for recognition, both in nature (for the attentions of natural selection) and by the apparatus of decision theory. We think that the birth of such an entity is a fundamentally important evolutionary innovation. It fills important gaps in the evolutionary history of cooperation.

Like Dawes, *et al.,* and in company with Bacharach, we have cast our construal of ecological challenges in terms of units of agency ("I" and "we"). Another way of framing the debate might be in terms of altruism. There has been considerable discussion of the cooperation question as a question of altruism versus self-interest, with altruism being cast as the antithesis to self-interest. We believe that the selfish/altruistic dichotomy is misguided because it elides the difference, on the nonselfish side, between acting as a member of a collective and acting in the interests of another entity, as an entity separate from it. In other words, the selfish/altruistic dichotomy neglects as distinct another and possibly quite potent and pervasive form of motivational organization: the individual acting as a part of a collectivity. When someone votes against their individual interest, either as a private citizen or as a member of a governing body, is that person performing an altruistic act? And when a person bothers to perform what they view as an admittedly minor civic duty, for example by exercising a legal right to vote, is that an act of altruism? We believe there is an important difference between *civic*-mindedness and bona fide *other*-mindedness. And this is the difference that our taxonomy of agentic forms attempts to capture. In our view, the most probing distinctions are made in terms of agency: X acts on behalf of Y, where X ranges over all entities that can serve as agents and Y ranges over all entities to whom an interest can be attributed.

When the capacity for bonding is impaired, the consequences are liable to be pervasive; both one's sense of community and one's sense of self suffer, because each (according to our proposal) is achieved through a very general capacity for bonding. With this conceptualization in place, it is easy to see how someone can be both notoriously intelligent and manipulative, and yet also imprudently impulsive. This description fits perfectly the classic characterization of the psychopath,[21] though the dichotomy between altruism and self-interest has made it tempting to downplay the psychopath's imprudent impulsiveness and highlight his indifference to others instead. With our construal of bonding as an elemental human capacity for both transacting and conceptualizing social reality, it becomes clear that the elements of the classic characterization of the psychopath are not in tension. Quite to the contrary, it is natural to find them going hand in hand. Insofar as

21] The modern conception of psychopathy was articulated by Hervey Cleckley in his classic work *The Mask of Sanity* (1941). According to Cleckley's criteria, a psychopath is intelligent, manipulative, irresponsible, impulsive, inadequately motivated and entirely devoid of shame.

psychopaths suffer a general deficit in the capacity for bonding, they are as prone to harm their future selves as they are to harm other living things.

VI. DIVISION OF LABOR WITHOUT HAPLODIPLOIDY: OFEK'S BAPTISM BY FIRE

Haim Ofek (2004) offers important new insights, and an ingenious argument for the evolution of division of labor without haplodiploidy. It has been customary for some time now to consider the prospects for cooperation furnished by the promise of so-called *public goods* – goods like protection from predators, which are such that if provided to some are effectively provided to all in the group (nonexclusion property), and whose enjoyment by one more does not detract from or lessen its enjoyment by the originals (nonrivalry property). Hence game-theoretical research has modeled evolutionary challenges as PDs, wherein individual interest enters into conflict with that of the group because no one individual has an individual motivation to contribute to the creation of the good if no one can be excluded from its enjoyment once it has been produced. This obstacle to cooperation is commonly known as free-ridership, which (as common wisdom has it) is responsible for underproduction of the good in question.

Ofek, by contrast, draws attention to the creation of *contrived goods* as providing the incentive for division of labor. He brilliantly discusses big game and the domestication of fire as instances of contrived goods. Contrived goods (like public goods) are nonrival, in that the enjoyment by one more does not detract from its enjoyment by the originals, but it differs importantly in the dimension of exclusion: a contrived good is something that can be withheld from a noncontributor. Thus there will be a strong incentive to "specialize" in production of those things that cost you next to nothing to produce for one more consumer on the margin, and which is of some considerable value to that next consumer provided he doesn't already have some of it, if he has something of value (to you) to exchange. In circumstances where exchange is possible and not (in itself) costly, division of labor will be more efficient. (We will discuss at further length below what efficiency amounts to.)

How, specifically, can fire take the form of a contrived good? First, it offers prospects of exclusion when its production involves investment or skill – actually a suite of three separate skills is required that can be mixed for efficiency depending upon availability of fuel and other resources – specifically: incendiary, containment and maintenance skills; but the limiting resource is ignition (chs. 9-10). As Ofek writes, "The enormity of this requirement is no longer fully appreciated by modern humans within easy reach of matches. But until a point not so distant in the past it still posed a major challenge to all fire users giving ample opportunity for exclusion" (151). Second, fire's capacity for self-generation offers its human handler the prospect of being able to make fire with fire at no extra cost.

Absent the ability to make fire on demand (which is arguably a much later development in evolution), humans had to make opportunistic use of what can be borrowed from

nature (from natural fires started by lightning, for example). And here now is Ofek's ingenious argument: once "borrowed", there are a number of scenarios for how to extend the use of this recruited resource. There are three likely scenarios. First is what Ofek calls the "campfire" scenario – the communal fire. Fires have an optimal size: too small and they die out easily, too big and they are fuel-inefficient and hard to contain. This poses a large question of social organization: a central fire open to all the elements has all the features of a public good, and so subject to free-ridership:

> However beneficial to society as a group, individuals willing to undertake the painstaking task of tending a continuously burning central fire, providing it with fuel, protecting it from the elements (and from human error) – strictly as a voluntary act – are not easy to come by… To rely on voluntary action for the purpose of the day-to-day provision of a routine service in the mundane arena of subsistence, is to expect slightly too much of the wrong species in a wrong setting. (159)

Second is the "private fire" scenario: continuously burning small fires maintained individually by "private" users, typically family groups. The problem with such a system – and this is now the key to Ofek's argument – is accommodation of the occasionally unlucky or undervigilant user whose fire dies out. Short of waiting for another wildfire, the unfortunate's solution would be to "borrow" from a neighbor who is more lucky or more vigilant. Donation to the unfortunate might at least initially seem a costless gesture to a desperate mendicant. And common knowledge that provision to the unfortunate is available might seem like cheap 'fire insurance" to all. But, or at any rate so Ofek argues, this situation is inherently unstable: why be vigilant in maintaining your own if borrowing fire is free and maintaining is expensive? This arrangement too, it seems, is subject to the miseries of free-ridership, so it cannot maintain itself:

> if everyone can borrow fire on demand, it is no longer in anybody's interest to undertake the painstaking task of maintaining an ongoing flame. If the players could not figure out their own interest for themselves in the short run (say, because they presumably lack rational behavior), natural selection would figure it out for them in the long run. Either way, one ends up with a system comprised exclusively of borrowers, no donors, and no fire. (160-1)

Ofek's preferred scenario is the "incendiary hub", which calls for a firekeeper (someone who specializes in fire maintenance and control, but who is rewarded for this service by those who can borrow from him or her). This is a much more efficient system: cutting the cost of fuel and labor, and if there is more than one hub, the firekeepers can provide insurance to each other either freely or at the price of a small favor to be returned. And the division of labor is off and running! An ingenious argument – but flawed. The defect lies with the key move vis-à-vis the instability of the "private fire" scenario. Effectively, this scenario is *not* doomed by free-ridership. Here is why. If I am a vigilant firekeeper, and others are not so vigilant, I will soon find myself more often donating than receiving donation. And I might be noticing that the beneficiaries of my generosity are faring better than I am. Does this incentivize me to ease up on my diligent efforts? By no means! Surely I

will instead find reason to redouble my efforts. With this will come an incentive to refuse handouts, or insist upon a reward for my services. If I go with the second option, the scenario becomes (more or less) Ofek's "incendiary hub" scenario. If I go instead with my first option – simply refusing to bestow favors (and instructing family to do likewise) – we have a new scenario: call it the "fire island" scenario, on which good firekeepers share only with family and friends. There is no incentive for free-ridership in this scenario. But it is no market scenario, nor is it unstable – indeed it is simply the condition of all private goods. It promotes division of labor or specialization no more than does any other advantage conferred by consumption of a private good. And so, while free-ridership dooms the prospects of free "fire insurance", it does not doom the private production and consumption of fire.

What remains of Ofek's conclusion? It has to be qualified as follows: If fire does not remain privately produced and consumed (confined to its islands), it can become a catalyst of market exchange (and the division of labor that comes with it). On this story, fire does not realize the explanatory promise of "sparking" a new era.

However we believe that Ofek's reasoning is much more promising than this bland conclusion would suggest. Less is more: if we leave out the instability conjecture, we are left with a better argument. Compare the three scenarios on offer: "campfire", "private fire" and "incendiary hub". "Campfire" suffers from free-ridership, so it is inherently unstable. "Private fire" will either devolve into "fire island" or it will evolve into "incendiary hub." "Fire island" involves underutilization of fire (it's inefficient); "incendiary hub" will produce the right amount of fire utilization and is considerably more efficient. Nature will choose "incendiary hub". And we can perhaps reiterate one of Ofek's earlier remarks: If the players could not figure out their own interest for themselves in the short run (due to cognitive limitations), natural selection would figure it out for them in the long run.

But what is nature figuring out? *Nature is figuring out which is the better (i.e. more efficient) COMMUNITY organizational strategy vis-à-vis fire production and utilization*. On this amended reasoning, selection is taking place at a level considerably higher than the organism level. And it is specifically favoring a certain agency structure, as such.

We have nowise, by focusing upon contrived goods, avoided the need to address this issue. Any more than focusing upon public goods was able to do so. (Ofek, it's worth mentioning, is open to multiple levels of natural selection, so might view our amendment as reasonably friendly.) But how did the group-level entities become eligible to compete in the evolutionary dance in the first place? It becomes clear that whether you prefer to focus upon public goods or contrived goods, you do not avoid the need to answer this question. And without an answer to this question, the evolutionary story is incomplete. Our bonding story is an attempt to address that question.

VII. MORAL EMOTIONS? TRIBAL INSTINCTS?

We have suggested that bonding can help account for the prevalence of human cooperation that characterizes the organizational structures in which we live. The two existing types of models of cooperation that come closest to our own are (1) models that put moral emotions at center stage, and (2) models that put tribal instincts at center stage.

Frank: the moral-emotions model of cross-lineage cooperation

In his work on cooperation, Robert Frank (2001) argues that "moral emotions," such as sympathy, can account for human cooperation. He points to the fact that physical proximity and communication, which promote the development of sympathetic bonds, are conducive to cooperation in PD situations. For Frank, moral emotions support cooperation because they have subtle physiological effects that function as hard-to-fake signals that would-be cooperators can rely upon.

We have two main qualms with Frank's position. The first is that it is based on the hasty assumption that the only sort of bonding that accounts for human cooperation is emotional bonding. Physical proximity and communication can promote the development not only of emotional bonds, but also of identification; and, like emotional attachment, identification is conducive to cooperation in PD situations. Moreover, phenomenologically, identification seems like a more plausible candidate than emotional attachment when it comes to accounting for cooperation between individuals who are more or less strangers and whose contact has been limited to a brief, unintense encounter, like a ten-minute discussion session.

Our second qualm with Frank's position is that it casts the connection between emotional attachment and cooperation as less direct than it seems to be, given that emotional attachment is a form of bonding. For Frank, moral emotions support cooperation as follows: they give rise to hard-to-fake physiological manifestations of moral emotion, which serve as reliable, observable signals of genuine commitment, which support mutual trust, which supports mutual aid. This way of casting the connection between emotional attachment and cooperation suggests that if we could not signal our commitment to one another, there would be no mutual aid. But this seems wrong. If I am emotionally attached to you and you are emotionally attached to me (perhaps because, a few years ago, we went through a series of emotionally charged experiences together), this bond will support mutual aid, even if we cannot signal our commitment to one another. While mutual awareness of our mutual attachment may reinforce our bond to one another, the bond itself is sufficient to account for mutual aid; mutual awareness of the bond via signaling is not essential. Similarly, if I identify with you and you identify with me (perhaps because we have something in common), this bond will support mutual aid, even if we cannot signal our commitment to one another. This is evidenced by the fact (mentioned above) that individuals tend to exhibit solidarity with respect to anonymous individuals whom they think of as in the same boat as themselves. Note that, as the last few sentences suggest,

conditions that promote one's bonding to another will often promote this other's bonding to oneself as well; so the frequency of *mutual* aid, as opposed to one-sided commitment, is easily comprehensible, even without appeal to hard-to-fake signals.

Richerson and Boyd: the tribal-instincts model of cross-lineage cooperation

According to Peter Richerson and Robert Boyd, we humans have a tribal instinct that "allows us to interact cooperatively with … a rather large set of distantly related or unrelated individuals with culturally defined boundaries" (1999, 255-6). This tribal instinct prompts us to divide ourselves into groups in accordance with a variety of markers, and to demonstrate commitment to group goals. The relevant groups are tribe-like in that they are relatively egalitarian, with charismatic leaders exerting influence but not authoritarian control. Like Frank, Richerson and Boyd (2001) see emotional attachment as crucial to cooperation. They maintain that

> commitments to group goals are deeply rooted in the emotions of individual humans who make up the groups. The threats and promises of leaders are only credible to the extent that followers will collectively back them up with passionate action. (87)

But in Richerson and Boyd's model it is emotional attachment to the group, rather than to individuals in the group, that is crucial. Richerson and Boyd explicitly make room for progroup commitments that are not the result of personal attachments (2001, 187-8 & 202-3).

In defense of their tribal instincts model, Richerson and Boyd effectively argue that commitment to group goals and altruistic sacrifice are greater in military groups that have a tribal structure (or nested tribal structure) than in military groups that attempt to function via straightforwardly authoritarian control. We do not want to deny that humans may have a tribal instinct, but there are, we believe, cases of cooperation between individuals who are more or less strangers that do not fit neatly within Richerson and Boyd's tribal instincts model. Consider the following case, which Richerson and Boyd bring up in defense of their view that progroup commitment need not rest on personal attachments:

> In his prototypical experiments, Tajfel (1981) told subjects that they were participating in a test of aesthetic judgment. They were shown pictures of paintings by Klee and Kandinsky and asked to indicate which they preferred. Subjects were then divided into two groups, supposedly on the basis of their aesthetic preferences, but in fact at random. The subjects' task was then to divide a sum of money among members of their own group or the other group. Subjects discriminated in favor of the sham in-group members (2001, 203).

In this case, there is no charismatic leader and no basis for the attribution of emotional attachment to the group. (Unlike the members of the successful military groups Richerson and Boyd focus on, the group members in Tajfel's experiments do not have anything like a history with their group, during which an emotional bond to their group could develop; nor have they any emotionally charged experiences with their group.) At least on the face of it, the bonding in Tajfel's experiments seems like a clear case of simple identification.

The subjects seem to be conceiving of themselves as a part of a "we" and acting accordingly, just as one might conceive of oneself as temporally extended and act accordingly.

Of course, if one assumes that emotional attachment is the only form of bonding, then all bonds (including *intra*personal bonds) will be interpreted as emotional bonds. It will then be tempting to downplay the phenomenology of emotion. This is not necessary if one recognizes that bonding is a genus.

VIII. OUR CONTRIBUTION: THE NATURAL HISTORY OF BONDING SCHEMES

In the magnificent *Guns, Germs and Steel,* Jared Diamond seeks to explain variations in the fates of human societies entirely by means of features of their biogeographical and ecological realities. Everything else – including all of culture and all sources of technological innovation and accumulation – he treats as sources of noise. Selection retains only what is of advantage to the organism/population in the particular niche it happens to occupy or help construct. One construal of the proposal is that it gives voice to the idea that ecological realities, when they speak, speak decisively; ecology is always the highest and most important driving force when it comes to evolution. This "ecology is destiny" idea lies behind the approach to the problem of explaining the prevalence of human large-scale living as a purely ecological/dynamical problem. It lies at the heart of the generalist objection we sought to disarm at the outset.

And herein also lies its weakness. The "ecology is destiny" thesis is rarely true, and when true, true only in the limit, true therefore only as an approximation. For in every instance the organism is always part of the "ecology" to which it must adapt. This can matter a great deal. In some cases (as in the case of modern humans) the organism and its conspecifics create or construct conditions – niches – that are totally unlike those in which their species originally appeared, and that figure prominently in its future evolution. Furthermore, an organism can enter into alliances with conspecifics, and indeed with members of different species, that shield it from harsh "ecology", whether constructed or simply found. These bonds are forged serendipitously, perhaps. But once forged, natural selection honors them. Natural selection does not dictate that each individual organism must go it alone against its "ecology." If that were the case, then organisms as such – as alliances among cells, many genetically unrelated – should never have arisen.[22] Natural selection acts on such entities – such agencies – as it encounters. It does not arbitrarily cast asunder what has bonded together; and so what has bonded together might co-evolve. For a mammal, bonding *becomes* a feature of ecology (a feature of what one has to contend with), because it is part of what its conspecifics do; and it is at one and the same time also a feature of what comes naturally to it, the organism, as well. For humans, bonding comes in more varieties, because of the human ability to conceptualize, but the different varieties

22] This point draws on argumentation in Sober and Wilson (1998) as well as in Okasha (2009).

have the same motivational effects. Bonding with nonkin is thus motivationally no different from emotional bonding.

And this, fundamentally, is why a generalist approach is insufficient: there must be room made for analyzing what the organism brings to its niche. And so a purely dynamical/ecological analysis of an evolutionary problem is insufficiently fine grained. The organism changes the specifics of the ecological problem that it finds, however minutely. And the effect can be amplified by many orders of magnitude, even exponentially, when the organism can ally itself with others to change the question of *who* the ecological problem is a problem *for.*

To handle what requires handling, attention must be given to the nature of bonding schemes, their emergence, and their contributions to – and interference with – realities that antecede them. For this analysis an exhaustive taxonomy of bonding schemes is required, as well as mechanisms that can give rise to them. We have taken a first, baby step in this direction, challenging the supposed platitude that the boundaries of agencies coincide exactly, always and everywhere, with the contours of an organism's skin (or fur or what-have-you). We are contending that the assumption to the effect that strategy selection happens, always and everywhere, at the level of the individual – which is implicit in so much evolutionary game theoretical analyses – is without foundation.[23]

It is important for those interested in modeling choice situations (in game theory, for example) to acknowledge units of agentic organization.[24] So to acknowledge is also to acknowledge a new task for decision theory to perform: decision theory needs to address the question of how entities navigate between acting as individuals and acting as members of larger units. The thing to keep in mind is that the unit of agency can vary and a methodological decision to treat, invariably, only one possible unit of agency as *the* unit of agency can result in serious distortions. And it can result in loss of a capacity to explain other, choice-related phenomena.

By suggesting that certain large-scale social behavior of humans rests on a general capacity for bonding with others, which can manifest in a variety of ways, we are resisting the image of choosing with which Frederick Schick opens his book – and which we quote at the foot of the title of this essay: "Life is a long trip in a cheap car," he writes. "In a dark country. Without a good map." We reply that perhaps once upon a time, in primordial days, long before our own species appeared on this terrestrial stage, life might have been as Schick describes. And indeed it might be that way for many organisms on the planet today. But the human species has never known life this way. Life for the modern *homo*

23] All of the research conducted by Skyrms and colleagues falls under this category. Moreover, while Skyrms applies algorithms that mimic natural selection, he by no means models the mechanisms of reproduction that can be thought of as mammalian; indeed, it is best to say that he does not model reproduction at all – reproduction is treated exogenously, and not linked to any processes in the model whatsoever (the most telling fact is that the "young" do not differ in any way from the "adult" organisms). It is most charitable therefore to say that Skyrms has not yet ventured into the area of multigenerational solutions.

24] Susan Hurley (1989) similarly urges a non-fixed conception of the units of agency.

sapiens may be (if she's especially lucky) a long trip, but thanks to human bonding capacities, there is at least one well-lit road for every dark one, public transportation is the rule rather than the exception, and likely as not there is someone willing to go with you some of the way.

This is all good news. The hard-headed "realist" who insists that humans are motivated entirely by personal or individual concerns is, by our lights, no true realist. While humans may not often be *true* altruists, they are nonetheless prepared to put out for groups of which they conceptualize themselves as members, however they achieve this identification. It takes really quite little to achieve this conceptualization, as we have discussed. And this is good news indeed, since the fate of the globe now depends heavily upon human cooperation across multiple boundaries.

m.thalos@utah.edu
c.andreou@utah.edu

REFERENCES

Axelrod, Robert. 1984. *Evolution of Cooperation*. New York: Basic Books.

Axelrod, Robert and William Hamilton. 1988. The evolution of cooperation. *Science* 242: 1385-90.

Bacharach, Michael. 1999. Interactive team reasoning: A contribution to the theory of cooperation. *Research in Economics* 53: 117-47.

———. 2006. *Beyond Individual Choice*. ed. Natalie Gold and Robert Sugden. Princeton, New Jersey: Princeton University Press.

Barkow, J., L. Cosmides, and J. Tooby, ed. 1995. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press.

Baron-Cohen, S. 1995. *Mindblindness*. Cambridge, MA: MIT/Bradford books.

Batterman, Robert. 1998. Game theoretic explanations and justice. *Philosophy of Science* 65: 76-102.

Binmore, Ken. 1994. *Playing Fair*. Cambridge, MA: MIT Press.

———. 1998. A utilitarian theory of political legitimacy. In *Economics, Values and Organization*, ed. Avner Ben-Ner and Louis Putterman. New York: Cambridge University Press.

Bolton, G. and A. Ockenfels. 2000. ERC – a theory of equity, reciprocity and competition. *American Economic Review* 90: 166-93.

Buss, David. 1994. *The Evolution of Desire: Strategies of Human Mating.* New York: Basic Books.

Buss, David and N. Malamuth, ed. 1996. *Sex, Power, Conflict: Evolutionary and Feminist Perspectives*. New York: Oxford University Press.

Byrne, Richard. 1995. *The Thinking Ape*. New York: Oxford University Press.

Byrne, R. W. and A. Whiten. 1985. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*. Oxford: Clarendon Press.

Carver, Charles and Michael Scheier. 2001. *On the Self-Regulation of Behavior*. New York: Cambridge University Press.

Cleckley, Hervey. 1941. *The Mask of Sanity*. Augusta,GA: Emily S. Cleckley.

Crawford, Charles and Dennis L. Krebs, ed. 1998. *Handbook of Evolutionary Psychology: Ideas, Issues and Applications.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Dawes, R. M., A. J. C. van de Kragt, and J. M. Orbell. 1997. Not me or thee but we: The importance of group identity in eliciting cooperation in dilemma situations: Experimental manipulations. In *Research on judgment and decision making*, ed. W. M. Goldstein and R. M. Hogarth. Cambridge: Cambridge University Press.

Demetriou, Andreas and Smaragda Kazi. 2001. *Unity and Modularity in the Mind and Self.* New York: Routledge.

Diamond, Jared. 1999. *Guns, Germs & Steel : The Fates of Human Societies*. London: W. W. Norton.

Duval, ThomasShelley, Paul J. Silvia and Neal Lalwani. 2001. *Self-awareness and Causal Attribution*. New York: Springer-Verlag.

Fehr, Ernst and K. M. Schmidt. 1999. A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114: 817-68.

Geanakoplos, John, David Pearce, and Ennio Stacchetti 1989. Psychological games and sequential rationality. *Games and Economic Behavior* 1: 6-79.

Fisher, R. A. 1958. *The Genetical Theory of Natural Selection*. NY: Dover.

Frank, Robert. 2001. Cooperation through emotional commitment. In *Evolution and the Capacity for Commitment*, ed. Randolph M. Nesse. New York: Russell Sage Foundation.

Gold, Natalie and Robert Sugden. 2007. Theories of team agency. In *Rationality and Commitment*, ed. Fabienne Peter and Hans Bernhard Schmid. Oxford : Oxford University Press.

Hargreaves-Heap, Shaun and Yanis Varoufakis. 2004. *Game Theory: A Critical Text*. London and New York: Routledge.

Harris, Paul. 1991. The work of the imagination. In *Natural Theories of* Mind, ed. Andrew Whiten. Cambridge: Basil Blackwell.

———. 1994. Understanding pretence. In Charles *Children's Early Understanding of Mind* Lewis and Peter Mitchell. New Jersey: Lawrence Erlbaum and Associates.

———. 1995. From simulation to folk psychology: The case for development. In *Folk Psychology: The Theory of Mind Debate*, ed. Martin Davies and Tony Stone, 207-231. New York: Blackwell.

Hobbes, Thomas. 1994 [1668]. *Leviathan.* ed. Edwin Curley. Cambridge: Hackett Press.

Hrdy, Sarah Blaffer. 1999. *Mother Nature*. New York: Ballantine Books.

———. 2009. *Mothers and Others.* New Haven: Harvard University Press.

Hurley, Susan. 1989. *Natural Reasons*. Oxford University Press

Kramer, R. M. and M. B. Brewer. 1986. Social group identity and the emergence of cooperation in resource conservation dilemmas. *Experimental social dilemmas,* ed. H. Wilke, D. Messick, and C. Rutte. Frankfurt am Main: Verlag Peter Lang.

Maynard-Smith. 1982. *Evolution and the Theory of Games*. New York: Cambridge University Press.

Nagel, Thomas. 1970. *The Possibility of Altruism*. Princeton: Princeton University Press.

Nisbett, Richard. 2003. *Geography of Thought: How Asians and Westerners Think Differently...and Why.* New York City: Free Press.

Ofek, Haim. 2004. *Second Nature: Economic origins of human evolution*. Cambridge University Press.

Okasha, S. 2009. *Evolution and the Levels of Selection.* New York: Oxford University Press.

Panksepp, Jaak. 2003. Feeling the pain of social loss. *Science* 302 (5643): 237-239.

Rabin, Matthew. 1993. Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281-302.

Richerson, Peter J. and Robert Boyd. 1985. *Culture and the Evolutionary Process.* University of Chicago Press.

———. 1999. Complex Societies: The evolutionary origins of a crude superorganism. *Human Nature* 10 (3), 253-89.
———. 2001. The evolution of subjective commitment to groups: A tribal instincts hypothesis. *Evolution and the Capacity for Commitment*, ed. Randolph M. Nesse. New York: Russell Sage Foundation.
Sally, D. 1995. Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society* 7: 58-92.
Schick, Frederick. 1997. *Making Choices*. New York: Cambridge University Press.
Selman, Robert. 1980. *The Growth of Interpersonal Understanding*. New York: Academic Press.
Skyrms, Brian. 1994. Sex and justice. *Journal of Philosophy* 91: 305-320.
———. 2003. *Stag Hunt and the Evolution of Social Structure*. New York: Cambridge University Press.
———. 1996. *Evolution of the Social Contract*. New York: Cambridge University Press.
Sober, E. and Wilson, D.S. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.
Sterelny, Kim. 2000. *The Evolution of Agency and Other Essays.* Cambridge: Cambridge University Press.
Strum, S.C., D. Forster and E. Hutchins. 1997. Why Machiavellian intelligence may not be Machiavellian. In *Machiavellian Intelligence II: Extensions and Evaluations*, ed. R. Byrne and A. Whiten. New York: Cambridge University Press.
Thalos, Mariam. 1999. Degrees of Freedom in the Social World. *Journal of Political Philosophy* 7: 453-77.
———. 2007. Sources of Behavior: Toward a Naturalistic, Control Metaphysics of Agency. *Distributed Cognition and the Will*, ed. Don Ross and David Spurrett, Harvard: MIT Press, 123-67.
———. 2008 . On Planning: Towards a Natural History of the Will. *Philosophical Papers* 37: 289-317.
Tomasello, Michael. 1999. *The Cultural Origins of Human Cognition*. Cambridge: Harvard University Press.
Wegner, Daniel and Robin Vallacher, 1980, eds. *Self in Social Psychology*. New York: Oxford University Press.
Wilson, E. O. 1975. *Sociobiology: The New Synthesis*. Cambridge: Harvard University Press.
Wilson, E.O. and Bert Hölldobler. 1990. *The Ants.* Cambridge, Mass: Harvard University Press.
———. 2009. *The Superorganism: The Beauty, Elegance, and Strangeness of Insect Societies.* New York: W.W. Norton & Company.

Public Reason 1 (2): 73-87

Political Realism and Political Idealism: The Difference that Evil Makes

Roman Altshuler
Stony Brook University

Abstract. According to a particular view of political realism, political expediency must always override moral considerations. Perhaps the strongest defense of such a theory is offered by Carl Schmitt in *The Concept of the Political*. A close examination of Schmitt's main presuppositions can therefore help to shed light on the tenuous relation between politics and morality. Schmitt's theory rests on two keystones. First, the political is seen as independent of and prior to morality. Second, genuine political theory depends on a view of human beings as evil by nature. I will argue that both claims are incomplete. Just as the political sometimes demands that morality be overridden, so morality can demand the overriding of political expediency. Moreover, the view of human beings as evil, which serves as the foundation of political realism, itself depends on affirming that human nature must also be, in some sense, good. Political realism is thus shown to have its theoretical foundation within a normative framework that demands the political pursuit of at least some moral aims.

Key words: evil, Kant, liberalism, political realism, Carl Schmitt.

In *The Concept of the Political*, Carl Schmitt presents a striking and highly powerful picture of the political. It is primarily characterized, on his view, by two important features. First, the concept of the political is succinctly defined in terms of an antithesis: the famous distinction of friend and enemy. Since this is seen as a *real* distinction as opposed to a normative one, this move allows Schmitt to argue for the independence of the political from all normative considerations and, in fact, to place it above them. Second, Schmitt separates what he thus presents as a political *realism* from normative political theories, particularly liberal ones, which we may contrast to the former by the designation of political *idealism*.[1] It follows from this brief sketch that theories of the latter kind are mistaken about the nature of the political. I will argue here that Schmitt's characterization of the political does not succeed. The first thesis about the political depends for its success on the second thesis about political theory. But the second thesis itself rests on a hidden normative premise about the nature of good and evil in human beings. I will conclude that the moral underpinnings of the argument invalidate the separation of political and moral spheres.

1] The term "political idealism" here is not meant to single out any particular political theory. Following Schmitt's own distinction, I use it to encompass all political theories that allow for normative components to play a guiding role in the political and, specifically, to do so *as* normative components. The latter qualification is needed for the following reason: Schmitt is willing to admit that normative considerations might play a role in the production of political conflicts or alliances – that is, moral or aesthetic agreements and disagreements might escalate to the political level. But they do not enter the domain of the political *as* normative agreements and disagreements; they merely provide material for conflict, which the political addresses without taking their normative status into account.

Both of these theses involve the relation between the political and the moral. I emphasize these particular aspects of Schmitt's thought because they are central to his political thinking as a whole, and also because I believe these to be the crucial points we must examine in order to avoid the trap Schmitt lays out for us, that is, the trap of too easily accepting the idea of the political as capable of being purified and separated from any relation to the moral. The place of morality in the two theses I will discuss is as follows. First, Schmitt defines the political by a criterion that distinguishes it and sets it apart from the spheres of economics, aesthetics, and morality.[2] Second, he insists that the view of human beings as evil rather than good by nature is the basic criterion by which *genuine* political theories may be distinguished from all others. In the first thesis, the relation between the political and the moral is overt; Schmitt explicitly presents the political as separate from the moral. In the second, however, the relation is covert, as Schmitt will attempt to de-moralize both good and evil insofar as they relate to the political. My goal will be primarily to demonstrate that Schmitt's reliance on this covert relation creates a conceptual difficulty for his political theory.

Although my argument here is largely a conceptual one, and aimed narrowly at a text no longer studied seriously by most political philosophers, I believe it has several wider repercussions. In particular, Schmitt provides us with one of the strongest arguments in the history of philosophy for the view of political considerations as surpassing and subsuming normative ones. What he defends, in other words, is *Realpolitik* in its strongest and barest form. Furthermore, he defends it on philosophical, rather than merely practical grounds. Schmitt's argument is not simply that expediency demands political decisions be made quickly and decisively without the delays and vacillations demanded by moral debate. For him, the demand to leave morality out of politics is not, at bottom, a matter of expediency at all. Rather, the nature of the political itself makes it immune to any moral considerations, and the importance of expediency is grounded in this nature. In today's political climate, which has barely emerged from a period in which justifications for war, torture, and suspension of the Geneva Conventions were presented as resting on considerations that trump moral concerns, Schmitt's argument is well worth revisiting.[3]

2] Schmitt sees the separation of the political from the economic, as well as the moral, as crucial to his case, because he takes the notion that economic influences are the driving force behind politics to be a dangerous dogma of liberal thought. On the other hand, the aesthetic threat is that the political may be transformed into spectacle (as in some postmodern theory or the media circus of American politics), obliterating the place of the political as an essential human concern. Here I will leave economics and aesthetics aside and focus instead on morality, since I believe the attempt to separate the political from normative concerns is the core of Schmitt's strategy.

3] In recent decades Schmitt's criticism of liberalism and democracy has also been appropriated, in interesting ways, by thinkers on the Left (see, e.g., Mouffe 1999). I only hint at some of the main challenges here, but it is worth noting that my argument does not curb the main thrust of these appropriated critiques.

I

The first thesis is a pivotal aspect of Schmitt's theory. In his attack on liberalism, he attempts to counter the liberal tendency to dispose of the political entirely, to reduce it simply to morality, to economics, or to aesthetic entertainment. Against a world that sees culture as consisting of mutually independent spheres outside the jurisdiction of the state, Schmitt posits the primacy of the political as if recalling a buried memory. By painstakingly tracing out the changes Schmitt made in the three successive editions of *The Concept of the Political*, Heinrich Meier (1995) demonstrates that the attempt to define the political grew radically more ambitious between the first edition of 1927 and the third in 1933. Schmitt began by attempting simply to procure an independent domain for the political against what he saw as a liberal tendency toward the depoliticization and neutralization of human existence. Initially, then, he wanted only to demonstrate that the political has its own sphere, distinct from other spheres of culture. Schmitt brings out the domain belonging to each sphere by positing an antithesis which defines that sphere: "Let us assume that in the domain of morality, the final distinction is good and evil; in aesthetics, beautiful and ugly; in economics, advantageous and disadvantageous or, for example, profitable and unprofitable." It follows that, to distinguish the political from these other spheres, Schmitt must formulate an antithetical pair as a criterion for it: "The specifically political distinction, to which political actions and motives refer, is the distinction between friend and enemy" (1963, 26).[4] Thus, all political activity is aimed at making a distinction between friends and enemies.

Whereas in 1927 Schmitt's goal of simply formulating and defending the distinctive features of a purely political domain seems fairly modest, by 1932 he expands his initial criticism of the liberal conception of autonomous domains of activity. The problem with this conception, for Schmitt, is that it covers over the political by attempting to shield all other domains of human culture from state interference. Liberalism is, thus, merely a bourgeois ideology intended to allow citizens to enjoy their wealth, their relations with others, and their entertainment activities in a purely private capacity. The point of breaking up culture into autonomous spheres is precisely to keep bourgeois activity out of the public domain and, consequently, away from the coercive power of the state. To counter this conception, it is not enough to simply present the political as one domain among others. It is crucial to show not only that it is independent of determination by any other sphere of culture, but also that it dominates over all other spheres. The political is now clearly not a sphere at all; it is, rather, a term for "the most extreme degree of intensity of a bond or a separation, of an association or a dissociation" (1963, 27).

4] The 1963 German reprint contains the text of the second edition, published by Schmitt in 1932. All translations from this work are my own, though I have benefited from George Schwab's efforts in Schmitt 1996. Readers interested in consulting Schwab's translation should note that the pagination is virtually identical to the German edition, for which reason I have not added it here.

The price of this expanded conception of the political is that Schmitt can no longer hold on to the idea of a "pure" politics, a political sphere analogous to the moral, aesthetic, and economic spheres. If the political refers only to the intensity of enmity, then this enmity must have its origin within some sphere or other that is not itself political. It may arise out of economic disagreements, moral disagreements, religious disagreements, and so on. Yet, although the political cannot exist without any other spheres, it remains pure in the sense that it is not dependent on any particular sphere. The political appears whenever a conflict can escalate to an extreme point, a point at which war becomes a possibility.[5] Schmitt's conception of the political now allows it to dominate over all domains of human activity in a way not possible so long as the political was only another sphere among others.

The argument for this primacy of the political is twofold. First, the political transcends all spheres of culture because each sphere may, in situations of conflict, escalate to the level at which it becomes politicized. Second, and more pertinent to my focus here, political concepts, because they involve the possibility of resolving a conflict through war, "have their real sense through the fact that they hold and maintain a reference to the real possibility of physical killing" (Schmitt 1963, 33). The emphasis on the word "real" here, shown through its repetition, is significant. The possibility of killing or of dying in war is a concrete possibility, not a theoretical or ideal one. This possibility also allows human beings to define themselves against an enemy. The participant in a conflict must "himself… decide whether the negation of his own kind of existence is signified in the otherness of the alien [i.e., the enemy] in the concrete, present case of conflict, and therefore whether that otherness will be fended off or fought in order to preserve one's own, proper kind of life" (27).[6]

The political attains a primacy, therefore, because it is the final forum in which human beings must identify themselves and define their "kind of existence" by being willing to kill or be killed to preserve it.[7] There are serious difficulties with this argument. First,

5] Of course Schmitt has already stacked the deck in his favor, so to speak, by insisting – without real argument or analysis – that each sphere of culture is defined by a central *opposition*. It is only a short step from this view of culture as an ongoing clash between diametrically opposed sides to the view that the political receives its mandate and its essence from the need to regulate and respond to these clashes.

6] The "himself" here is misleading, because Schmitt does not mean that each individual is the proper judge of friend and enemy. Although of course individuals can decide who their friends and enemies are, at the political level such decisions can only be made by the sovereign. Schmitt's wording here is particularly confusing since he is at pains to separate political conflicts from individuals ones, and the political notion of friend and enemy (in which the enemy is always a "public enemy") from the one contained in the Christian injunction to love one's enemy (29-30).

7] C.f. J. S. Mill's oft-quoted lines regarding war: "War is an ugly thing, but not the ugliest of things: the decayed and degraded state of moral and patriotic feeling which thinks nothing *worth* a war, is worse… A man who has nothing which he is willing to fight for, nothing which he cares more about than he does about his personal safety, is a miserable creature who has no chance of being free, unless made and kept so by the exertions of better men than himself. As long as justice and injustice have not terminated *their* ever-renewing fight for ascendancy in the affairs of mankind, human beings must be willing, when need is, to do

Schmitt's view of the political involves one's self-definition at a group level, while the moral involves individual convictions and individual identity. That Schmitt places collective self-definition above personal self-definition as part of his argument for the primacy of the political is at best question-begging, but it is problematic on other grounds as well. The group identity in question here is explicitly affirmed in the political, but this can happen only because a group identity is already a prerequisite for any political conflict. In other words, Schmitt assumes the existence of nations involving a common identity among members as a foundation of the political, thereby excluding as anti-political any grouping that seeks to establish (rather than defend) its identity through political processes.[8] The distinction between collective political action and individual moral action provides Schmitt with a convenient way of avoiding one kind of conflict between morality and politics. On the one hand, morality can require individuals to die in opposition to unjust laws, that is, in battle against the political. On the other hand, individuals may be required to kill others by political considerations, thereby violating their moral obligations. Schmitt addresses only this latter conflict, as it takes pride of place in his attempt to purify the political of any moral normativity.

But, and closer to the point I want to press later against Schmitt, the emphasis on death provides an odd foundation for an argument defending the primacy of the political over the moral. After all, morality has traditionally appeared with reference to its own extreme possibility, i.e., the possibility of dying for one's moral convictions. We find very clear arguments to this effect in, for example, Socrates's *Phaedo* and *Crito*, Augustine's *On the Free Choice of the Will* or in Kant's *Critique of Practical Reason*. If we attempt to articulate the distinction between the possibility of death in moral discourse and its possibility in the political domain, we arrive at two avenues.

A. The political, more so than the moral, involves not simply the possibility of dying, but also that of killing. Schmitt's argument is not merely that the political requires physical killing of one's enemy in conflict with morality, but rather that morality has nothing to say about this at all. In an extreme case of conflict, one is called on by the political to kill one's enemy, but "there is no rational end, no norm no matter how right, no program no matter how exemplary, no social ideal no matter how beautiful, no legitimacy or legality, that could justify human beings' killing each other for it" (Schmitt 1963, 49-50). The situation here is quite dire: killing one's enemy may be a political necessity, yet a necessity that cannot be justified by any moral norm. If we accept this view of morality, there appears to be only one way in which the moralist or the political idealist can respond. One can argue simply that nothing more is proven here than that the political as such is immoral. If the waging of war cannot be morally justified – if it is, in fact, immoral – then the domain of the political is itself evil. Schmitt's response to such an objection is that it is easily trumped

battle for the one against the other." (1862, 683-84)

8] For a development of this criticism, see Habermas 1998, 141-42.

by political realism. The idealist, in claiming that war is evil and refusing to go beyond this claim, will be unable to respond adequately to the real conflicts that materialize within the concrete practice of politics. An idealist will either be forced to avoid war at any cost, or will have to find moral justifications for war. The principled avoidance of war is, however, unfeasible: in cases of intense conflict, it would lead to the annihilation of one's form of life by an opposing enemy, so that the way of life that makes the principled avoidance possible in the first place would be lost. If fighting off an enemy becomes necessary, then, the idealist will resort to a moral justification, which will present the enemy as evil, inhuman, and deserving of being killed. But the problem with this approach is precisely that, in attempting to encompass the political, morality corrupts itself. By seeking to absorb the political, morality becomes its handmaiden, merely another tool of political discourse.

We seem, then, to be led to the following conclusion: if morality cannot justify political activity, and if that activity is necessary to human existence, then the political must be entirely immune from moral criticism. A decision to kill the enemy in order to preserve one's form of life is neither moral nor immoral. It is simply political. The independence of the political from the moral thus appears established. Yet this does not guarantee the *priority* of the political. For, even if the political is victorious in this conflict over the immorality (or, here, amorality) of killing, it is not yet clear why the moral may not likewise come out ahead in the other extreme conflict, where one is morally obligated to die in order to oppose political decisions. In other words, it still seems possible that, while the political is independent of morality, morality may likewise be independent of the political and may still be prior to it.[9]

That Schmitt does not confront this challenge head on may be due, perhaps, to an intentional oversight on his part. In rejecting the liberal view of autonomous, independent spheres of human activity, he is concerned only with establishing that these spheres are not independent of the political. He does not, however, address the possibility that the spheres are not at all independent of each other; in fact, Schmitt explicitly wishes to maintain that they *are* autonomous with regard to each other, since this allows him to place all human activity under the reign of the political without requiring a similar contamination in return; and it is a crucial part of Schmitt's critique of liberalism that the political be understood on its own terms, apart from determination by economic, moral, or other criteria. Thus, Schmitt places morality within its own sphere of bourgeois life. "The individual may voluntarily die for whatever he wants to; that is, like everything essential in an individualistic-liberal society, for all intents and purposes a 'private matter'" (Schmitt 1963, 49). The decision to die for a moral belief seems, to say the least, to be an odd thing to place under the category of a comfortable bourgeois existence.

9] Of course morality would only be prior from the moral perspective, or the perspective of a conflict described as a moral one. The political would have priority from the political perspective. My point, however, is that in insisting on a radical separation between the two domains, Schmitt begs the question against the moral perspective.

Schmitt's view that politics is destiny[10] seems to rest precisely on this distinction between public and private. The individual may be forced to kill, despite his or her own moral objections, in order to preserve a public existence. The decision to make a group of people kill another group is a concrete possibility, not a private matter. That is, the individual cannot simply shrug off such a requirement, but is necessarily confronted with it. A moral decision, however, *is* a private decision, and the individual may choose to ignore it at will. A citizen called to fight in a war may choose not to heed the call and to face the consequences. But he or she is not forced to make such a choice on moral grounds. Human beings are faced with the concrete reality of the political in a way that they are not faced with the reality of the moral. But here the political idealist has a clear rejoinder: that the individual is not forced to make a moral choice is the wrong conclusion to make. Rather, in a society that has forgotten the importance of morality (just as, according to Schmitt, liberal society has forgotten the importance of the political), the individual can ignore the fact that his or her decision is a moral one. That one is not explicitly faced with a moral decision *as* a moral decision does not mean that the moral decision is not a concrete one. Historically, this view is clear in the consensus at the Nuremburg trials that "I was only following orders" does not excuse one from the charge of crimes against humanity.

B. The decision to compel the citizens of a state to kill is made by the sovereign; if anything, then, and in opposition to Schmitt's own expressed political leanings, one may suggest that this is precisely a key argument in defense of a (liberal) republican, rather than authoritarian, political system. This, in fact, is Kant's argument in defense of his "First Definitive Article for Perpetual Peace": the more the sovereignty is in the hands of those who are forced to suffer losses in war, the less the likelihood of conflict escalating to such a level (Kant 1999, Ak. 8:349-50).[11]

I would suggest, then, that Schmitt succeeds in demonstrating the autonomy of the political (with regard to morality) *according to his conception of the political*; the question of the *primacy* of the political, however, remains unresolved. That the political has a concrete reality and, consequently, a primacy over the moral remains an assertion based on the merely *apparent* relative degree of reality of the political over the moral. Ultimately, however, both the question of the independence of the political from the moral, and that of its primacy, rest on the question of whether human beings are, really, moral beings in any significant sense. Here we move to the second aspect of Schmitt's thought that I wish to address.

10] This claim is attributed to Schmitt in a powerful summary of his view: "The political is a basic characteristic of human life; politics in this sense *is* destiny; therefore man cannot escape politics" (Strauss 1932, 94).

11] All references to Kant will give the volume and page number from the German Akademie edition of *Kants gesammelte Schriften*, standardly used in writing on Kant and available in the margins of most recent English translations of Kant's works.

II

Schmitt's criterion for the classification of political theories is the following: "One could examine all theories of the State and political ideas according to their anthropology, and divide them based on whether they, consciously or unconsciously, presuppose man to be 'by nature evil' or 'by nature good'" (1963, 59). At first glance, of course, this seems to suggest that the division of political theories is, in fact, grounded in the same antithesis of good and evil that provides the criterion for the moral sphere. That impression is, however, misleading, as Schmitt goes on to state that, "The distinction is entirely summary and not to be taken in any special moral or ethical sense. Decisive is the problematic or unproblematic nature of man as the presupposition of every further political consideration, the answer to the question whether man is a 'dangerous' or not dangerous, risky or harmless and not risky being" (1963, 59). "Good" and "evil" are not, here, moral terms; they are, rather, anthropological ones, as indicated already by the "by nature" prefixed to each term. Whether or not this de-moralization of moral concepts can be made coherent is a question I will return to momentarily. First, however, it is important to draw attention to the way in which this distinction is from the outset quite different from the moral one.

"All genuine political theories presuppose man to be 'evil,' viz., consider him not at all as an unproblematic, but rather as a 'dangerous' and dynamic being" (Schmitt 1963, 61). This evaluative classification follows from the main lines of Schmitt's theory that have already been sketched out. The liberal or political idealist sees human beings as good by nature; such a thinker therefore either refuses to accept the possibility of having an enemy who must be killed or invents an excuse that makes the enemy out to be evil, either because that enemy is somehow less than human, or because he has deviated from natural goodness. The political realist – or the genuine political thinker, in Schmitt's terminology – accepts human evil; thus, he or she requires no *moral* justification when the necessity for war arises. Furthermore, the believer in human good ultimately doubts the need for government; if human beings are truly good, then government may be superfluous; it may, perhaps, have no important reason for being other than to regulate commerce and deal with the occasional deviant, as one reading of Locke's *Second Treatise* suggests. A thinker who believes that human beings are evil, on the other hand, recognizes the eternal need for a government that accepts the possible necessity for war and can decide on the basis of each case whether such a necessity has arisen. A political theorist who views human beings as good, then, cannot be a *genuine* political thinker, since the presupposition of human goodness, by undermining the need for government, undermines the necessity of political thought. Schmitt thus suggests that, taken to its extreme conclusion, the distintion between good and evil provides the basis for the corresponding political distinction between authoritarians and anarchists.[12]

12] This reduction of the main lines of political thought to either authoritarian or anarchist is, of course, intended to echo Hobbes's derivation of the rationality of indivisible and absolute sovereignty from

The important point to note is that the antithesis of good and evil does not function here in the same way that Schmitt sees it functioning in the moral sphere. For *here* only one side, the side of evil, is identified with a genuine theory. The side of good is misguided, or worse. Schmitt does not simply prioritize evil; he puts in question the reality of the good. From a *genuinely* political perspective, a perspective of realism, the good is only a fiction. Evil belongs to human nature, to concrete human reality.

To examine this point, we may return to Schmitt's earlier claim that morality is distinguished by the antithesis of good and evil. This is a dubious characterization of morality, at least in historical terms, as there is a strong historical tradition, stretching from Socrates though Augustine and beyond, which denies the reality of evil altogether. According to this tradition, which we may here label as the tradition of transcendental morality (TM), evil is simply the absence or privation of good. Despite introducing further complexity into the analysis, Kant is perhaps the most prominent modern defender of this tradition. In a passage in his *Religion Within the Limits of Reason Alone*, Kant criticizes the Stoics (and, by extension, the entire tradition) for failing to recognize that the enemy of good is not simply a lack of discipline or wisdom, but is a real enemy: the principle of evil (Kant 1960, Ak. 6:57). Yet Kant's criticism here conceals his own bias in the presentation of the antithesis of good and evil and his underlying sympathies with the tradition. Both good and evil, for Kant, are defined in relation to the moral law. But evil is defined negatively, as a principle of occasional deviation from the law, while good is derived from the law itself. Furthermore, it is only because we have the moral law within us pointing us to the good that it makes sense to hold us morally accountable for our actions; and thus, again, it is only in relation to our possession of the law that we can be evil at all. Finally, the law is a product of reason, which regulates our understanding of nature. Deviation from the law, on the other hand, is occasioned by our surrender to natural inclinations. Although Kant asserts that human evil is universal, it remains the case that the good is metaphysically prior to evil; it is the standard against which evil is defined. Good and evil do not, therefore, form a true antithesis within the tradition of TM. If we limit our analysis of morality to TM,[13] then it looks like the political, as Schmitt has described it, is in fact the flip side of the moral. The political presents evil as the defining term while denying the concrete reality of good. The moral, on the other hand, allows for the possibility of evil only through mediation by the good. To clarify, we can draw out the contrast – as well as the analogy – between morality, as conceived by TM, and politics, in Schmitt's political realist account:

---

the conditions of the state of nature – a derivation, as Schmitt is clearly aware, grounded on a presumption of human evil.

13] This is not the place to make this argument, but I believe that TM provides the strongest account of the sort of morality capable of trumping one's inclinations, including self-interested motives and, in extreme cases, the instinct of self-preservation. In any case, some version of TM remains common in moral theory, especially in Kantian and (some) Aristotelian approaches.

TM: Morality presupposes the reality of the good as an ideal inherent in human nature, such that evil is understood merely as the absence of good or a failure to live up to the ideal.

PR: Political theory presupposes the reality of evil. Attributing reality (within human nature) to the good misrepresents concrete political reality and undermines the possibility of genuine political theory.

The appearance of a contrast may not be entirely accurate, however, since Schmitt has claimed that good and evil are not, in his differentiation of political systems, to be taken in a moral sense. How, then, should we take them? In discussing the variations that these concepts undergo in political anthropologies, Schmitt offers the following ones: "'Evil' can appear as corruption, weakness, cowardice, stupidity or also as 'beastliness,' instinctual drivenness, vitality, irrationality, etc., the 'good' in corresponding variations as rationality, perfectibility, tractability, educatability, congenial peacefulness, etc." (1963, 59).

Leo Strauss, in his "Notes on Carl Schmitt", argued persuasively that this distinction is untenable. The first set of definitions of evil presents it already in a moral light, and thus must be ruled out by Schmitt's criteria. The second set, on the other hand, involves an "innocent" evil, an evil belonging simply to animality. But if human beings are evil merely by virtue of being animals, then they can be trained; and if they can be trained, then they can be educated. But "the limits one sets for education finally become a matter of mere '*supposition*' – whether very narrow limits, as set by Hobbes himself, who therefore became an adherent of absolute monarchy; or broader limits such as those of liberalism; or whether one imagines education as capable of just about everything, as anarchism does" (Strauss 1932, 99). Since educatability falls, for Schmitt, on the side of good, the distinction between good and evil collapses.[14]

Strauss does not, however, address the list of variations for "good." We may note from the outset that "corruptibility" must belong on this side, just as "corruption" belongs on the side of evil. A *genuine* political thinker views human beings as naturally corrupt; a misguided thinker, of course, in seeing human beings as good, must view corruption as

14] As Meier convincingly demonstrates by examining the changes Schmitt made in his third edition in response to Strauss's critique, his real notion of evil has little to do with animality and much to do with original sin. If the political may rely on the notion of original sin, however, it is not entirely clear what justifies the denial of divine grace. But I will not pursue this issue here, as my concern is with the philosophical, not the theological implications of Schmitt's thought. Schmitt scholars may respond that in ignoring the theological implications I am, in effect, ignoring Schmitt's own commitments. Since I am interested here primarily in the implications of Schmitt's thought for the relation between politics and morality within the framework of secular theory, however, I propose pursuing the issue within that arena, leaving the theological questions for work in the domain of Schmitt studies. Since a radical separation between politics and morality is a threat in secular, as well as theological, circles, and since Schmitt on my view provides a powerful defense of that separation, I hope my lack of attention to the theological details is justified for the purpose at hand.

a fall from innocence. What Schmitt has in mind when referring to theorists on the side of good, I suspect, are thinkers like Rousseau, as well as Kant, whose view can be summarized by his interpretation of the biblical account of the fall: "man is represented as having fallen into evil only *through seduction*, and hence as being *not basically* corrupt… but rather as still capable of an improvement" (Kant 1960, Ak. 6:44). Even Nietzsche fits on this side, contrary to Schmitt's protestations, for despite his sympathies with evil, he insist that Christian morality can and in fact has tamed humanity; the appearance of so-called "higher men" who free themselves from that yoke can, from this perspective, be seen as a rare aberration that poses little threat to the moral order.[15] The theorist of evil, by contrast, sees human beings as always corrupt and therefore as incapable of improvement. Leaving aside the rhetorical question of how many thinkers have, in fact, shared such a strict view, we may turn to what is entailed by it.

First, Schmitt's suggestion that the genuine political theorist views human beings as dangerous and *dynamic* is questionable. A dynamic being, one would think, is capable of change; it is a being that can either be improved from a corrupt to an uncorrupted state, or vice versa. But both of those views, the possibility of being improved (educatability) and the possibility of being corrupted (which implies a natural lack of corruption), fall squarely on the side of good. What is left for the side of evil, in fact, is an entirely static being; a being that is dangerous not because it is dynamic, but because it will predictably shrug off every norm when given a chance. And this is, in fact, what seems required for Schmitt to associate the side of evil with genuine political thought. After all, if human beings are capable of improvement, then the perpetual need for government, for the political itself, is undermined. Either it was never needed to start with, or it can disappear once humanity has reached the proper level of improvement. The static nature of evil gives it the appearance of concreteness, of reality. Unlike the good, which is grounded in conjectures about improvement, evil can be presented as a brute fact.

A second look at the list of variations may suggest that the side of good is (in TM terminology) full of capacities, while the side of evil is largely populated by incapacities. The capacities to be educated, to become tractable or capable of perfection, are genuinely *abilities*, belonging to human goodness. On the side of evil is perpetual corruption, drivenness by instinct, irrationality: these imply a lack, an inability to change or to be changed. What distinguishes evil, what sets it apart, is precisely its lack of dynamism: it is not *becoming*; it is concrete, actual *being*. Change, for better or worse, is an unpredictable, unclear ideal. Sameness is true reality. Remaining the same, refusing – not by choice, but *by nature* – to be tamed or educated, involves an inability to adapt and adjust to new circumstances. And this is precisely Schmitt's underlying point: it is *because* human beings do not and cannot change, from his perspective, that the political always remains necessary; changeable be-

15] That, of course, is not Nietzsche's own view. The point is only that his account of the transformation of human nature through Christianity assumes a tamable and educatable human nature, that is, a nature "good" in Schmitt's sense.

ings might change their destiny, they might outgrow the political, but for human beings who are evil by nature, politics is destiny.

If Schmitt succeeds in demoralizing moral terminology at all, he succeeds only on the side of evil. Good, insofar as it may involve the ability to be educated and tractable, allows for the human possibility of being governed by moral norms. Evil beings, on the other hand, require the constant presence of the watchful eye of the political. They require it because moral norms will not stick to them, because their unchangeability makes them eternally dangerous. But this distinction between a demoralized, "anthropological," factual notion of evil and a moralized, or at least always potentially moralized, notion of good seems, once again, analogous to the TM tradition.[16] And once again, following through with the analogy, we may ask: what does this reality, this concrete unchangeability, mean at all? Where does it get its supposed reality? I propose that it is precisely by contrast with the good, with the moral, idealized, changeable notion of humanity. Insofar as the good is seen as merely ideal, its opposite can be presented as real. My argument does not require going as far as the position of TM: it is not necessary to insist, dogmatically, that evil can be defined only through the priority of the good. My point, rather, is that evil, even a demoralized, concrete, *real* evil, is incomprehensible without the possibility of the good, understood in a moral sense.

So what does the task of demoralizing evil amount to? As noted above, Strauss argues in his critique that "the opposition between evil and good loses its keen edge, it loses its very meaning, as soon as evil is understood as innocent 'evil' and thereby goodness is understood as an aspect of evil itself." And Strauss goes on to suggest that Schmitt's aim is not to demoralize evil at all, but to affirm it, in a moral sense, in order to affirm the political. "The task therefore arises – for purposes of the radical critique of liberalism that Schmitt strives for – of nullifying the view of human evil as animal and thus innocent evil, and to return to the view of human evil as moral baseness" (1932, 99). A very natural reading of Schmitt, one according to which his aim is really a politically oriented revaluation of morals, arises from this trajectory.

Instead of following Strauss's path, I have attempted to trace the logic that a genuinely demoralized evil demands. As I have indicated, tracing this path is crucial if we are to understand political realism, since if political realism merely affirms a moral evil, then it is simply immoral. The stronger goal of giving the political a priority over and independence from the moral requires that morality be excised from anthropology altogether. But as I

16] The reference to possibility or potential here is important, especially in showing why the notion of a good in human nature tends toward a specifically *moral* good, rather than simply an innocent animal good. Among the human capacities is rationality, and this means that unlike the other animals, human beings who are capable of being educated are also capable of following their reason and, in Kantian terminology, acting on their conception of a law and thus subjecting themselves to norms. As Schmitt's reference to evil as irrationality suggests, his anthropology insists not that human beings lack reason, but that they do not normally follow it. But this view is commonly upheld by the TM tradition, whose defenders note that it is only the *capacity* for reason that is needed to hold us accountable for our moral failings, but also to allow us to strive to overcome them. Human beings may all be evil, but they are also all potentially good.

have been attempting to show, this project involves taking the TM view of human nature and cutting out the ideal of the good along with any human capacities conducive to striving for it. What is left – demoralized evil – can then be presented as pure, unchangeable, and real. The move is not from a view of human nature to a view of the political; rather, political realism begins with the concept of the political and then affirms the anthropology presupposed by it. And this means that the argument can just as easily be reversed. If we start with a more complete anthropology, it turns out that while Schmitt's concept of the political may be correct, it is also grossly incomplete.

III

Let us attempt, in conclusion, to find the key to Schmitt's theory. We can state it as a paraphrase of Schmitt's claim: Any political realism presupposes that man is by nature evil. That is, any theory that assumes that the political deals exclusively with real and concrete conflicts, that defines the normal functioning of the political – as well as the necessity of such functioning – in relation to exceptional situations, takes as its presupposition the view of humanity as evil. If human beings are irredeemably evil, the political must occupy itself entirely with a behavior oriented toward the possibility of war. But there is a flip-side: insofar as the human being is anything other than that; insofar as there is any potential for something beyond stupid, corrupt dangerousness, the political cannot be defined exclusively by the antithesis of friend and enemy.

And here is the crux: Even if we accept the real need for such distinctions, and even if we grant that behavior oriented toward the possibility of war is at least partially motivated independently of moral distinctions, two outstanding claims of Schmitt's political realism remain to be substantiated. The first is the claim which we have already examined: that of the autonomy of the political. If the political can take precedence over the moral, it remains unclear why, on either a practical or a theoretical view, a reverse contamination cannot occur. Second, the dominance of the political is put into question; for if the moral can enter into the sphere of the political, there remains always the possibility that the political may – if human beings are even potentially capable of improvement – similarly come to be dominated by the moral.

But why are we challenging the claim, which earlier seemed solid, that the political is autonomous with regard to the moral? The answer, already hinted at, is that a politics that concerns itself exclusively with intense association and dissociation is half-blind: it misses the potential for good, which is necessary for a proper understanding of evil. My suggestion is not that the grouping of humanity into friends and enemies is not a concern for the political; nor am I arguing that such activity does not enjoy some autonomy from the moral, although I suspect such an argument could be made. Rather, the point is that this cannot be the *exclusive* concern of politics. The political cannot be fully outside the moral realm, the realm of human improvability. If human beings have a capacity for betterment, the political cannot simply ignore that capacity while remaining morally neutral.

To quote Kant one more time, "moralizing politicians, by glossing over political principles contrary to right on the pretext that human nature is not *capable* of what is good in accord with that idea, as reason prescribes it, *make* improvement *impossible* and perpetuate, as far as they can, violations of right" (1999, Ak 8:373). Politics cannot act exclusively on the assumption of human evil without itself becoming evil, and evil in a moral rather than anthropological sense.

That the political has a role to play in the *development* of human beings, and not simply in the possibility of their killing each other, may already be gathered from an analysis of Schmitt's own remarks concerning the potential need to fight off an enemy "in order to preserve one's own, proper kind of life" (1963, 27). Already implied in this statement is the need for *preservation* of a kind of life, a preservation that simply cannot occur exclusively through an encounter with an enemy. The political must have, as one of its functions, the nurturing of the potential for good in human nature; if this function does not necessarily overrule the conflictual function of the political based on human evil, it must at least complement that function. If human beings are both evil and good, in some sense, the political must address itself to both sides of our anthropology.

Preservation is, of course, an ambiguous term. It could mean nothing more than maintaining the status quo. But if human beings are genuinely incapable of change, it is unclear why preserving any particular way of life over another might be worthwhile: any way of life is just as good, or just as bad, as another. To be worthwhile, then, it seems preservation must involve something more: the preservation of a way of life not only at an existing stage of development, but rather the preservation of this development itself. As a function of the political, the grouping of humanity into friends and enemies must be oriented toward the purpose of nurturing or at least allowing for improvement, "for whatever might be the highest degree of perfection at which humanity must stop, and however great a gulf must remain between the idea and its execution, no one can or should try to determine this, just because it is freedom that can go beyond every proposed boundary" (Kant 1998, A317/B374).

To place the political entirely beyond the reach of moral criticism, then, is to leave it aimless and blind. Nor should we be satisfied only with meeting Schmitt half-way by suggesting, for example, that the political may have two aspects, a real and an ideal one, that do not intersect. It is one thing to insist that the political requires expediency, that it is meaningless without the possibility of war, and even that the need for war may not rest entirely on the moral for its legitimacy, so that some political decisions may appear both necessary and immoral. But it is another thing to insist that the political can operate entirely without serving moral purposes or that it can be fully immune from moral criticism. Balancing genuinely moral with genuinely political considerations – if we can still meaningfully make the distinction – is a difficult task, and one that frequently threatens to undermine the grounds of both sides involved. But this balancing act is poorly described

as involving a conflict between independent spheres, or between collective and merely private considerations. It is a conflict internal to political theory itself.[17]

This conclusion may be prosaic. But my point is precisely that Schmitt's analysis of the political appears so exciting *because* it obscures prosaic truths. It is captivating because it seems to make feasible what is really unfeasible: the possibility of a coherent notion of evil without a notion of good. And it is only if we first accept such a possibility that we can allow for the more serious possibility that Schmitt pushes on us, that is, the possibility of a politics purified of moral content.[18]

raltshul@ic.sunysb.edu

REFERENCES

Habermas, Jürgen. 1998. *The Inclusion of the Other*. Ed. Ciaran Cronin and Pablo De Greiff. Cambridge, MA: MIT Press.

Kant, Immanuel. 1998. *Critique of Pure Reason*. Trans. Paul Guyer and Alan Wood. New York: Cambridge University Press.

———. 1999. Toward Perpetual Peace. Reprinted in Mary J. Gregor, Trans. *Practical Philosophy*. New York: Cambridge University Press.

———. 1960. *Religion Within the Limits of Reason Alone*. Trans. Theodore Greene and Hoyt Hudson. New York: Harper & Row.

Meier, Heinrich. 1995. *Carl Schmitt & Leo Strauss: The Hidden Dialogue*. Trans. J. Harvey Lomax. Chicago: University of Chicago Press.

Mill, John Stuart. 1862. The Contest in America. *Harper's New Monthly Magazine*, April.

Mouffe, Chantal, ed. 1999. *The Challenge of Carl Schmitt*. New York: Verso.

Schmitt, Carl. 1996. *The Concept of the Political*. Trans. George Schwab. Chicago: University of Chicago Press.

———. 2002 (1963). *Der Begriff des Politischen*. Berlin: Duncker & Humblot.

Strauss, Leo. 1996 (1932). Notes on Carl Schmitt, *The Concept of the Political*. Trans. J. Harvey Lomax. Reprinted in Schmitt 1996.

17] And, it almost goes without saying, to moral theory itself.

18] I would like to thank Prof. Dr. László Tengelyi and other audience members at my panel at the International Association for Philosophy and Literature in Freiburg, June 2006, for valuable comments on an earlier draft of this paper.

Public Reason 1 (2): 88-93

How the Ceteris Paribus Principles of Morality Lie

Peter Shiu-Hwa Tsu
Australian National University

Abstract. This paper addresses the issue of how the ceteris paribus principles of morality lie. I discuss a recent attempt by Margaret Little and Mark Lance to cash out ceteris paribus principles of morality in terms of "nature" and "privileged conditions". I argue that when the ceteris paribus principles of morality are so cashed out, they cannot be plausibly claimed to be true.

Key words: Margaret Little, Mark Lance, particularism, defeasibility, ceteris paribus.

In their recent co-authored articles, "Defending Moral Particularism" (2006) and "From Particularism to Defeasibility in Ethics" (2008), Margaret Little and Mark Lance argue for the truth of ceteris paribus moral principles and their indispensability for morality. In one of the passages, they contend that to say that "ceteris paribus, lying is wrong" is in essence "equivalent to asserting an *exceptionless* connection between lying and a milder moral property: lying may sometimes be morally justified, but it is *always* wrong-making" (2006, 306; emphases added). In another, they contend that "it is important to any adequate morality to recognize that [ceteris paribus], pain is [bad]" (2008, 68; all later paginations refer to Little and Lance's 2008 article). I take it from these passages that they argue for the following two theses: (1) **The Thesis of Truth**: ceteris paribus moral principles are true in the sense that they are exceptionless; ceteris paribus, lying is always wrong; or equivalently, lying is always wrong-making. (2) **The Thesis of Recognition**: Any adequate morality has to recognize the truth of ceteris paribus moral principles. In this paper, I focus mostly on the assessment of (1) and argue that Little and Lance do not provide compelling arguments for us to accept it. If I am right about this, some doubts can be cast on (2), for there are no compelling reasons to believe that there are true ceteris paribus moral principles to be recognized.

I. TWO GENERAL CHARACTERISTICS

In order for us to determine whether a ceteris paribus moral principle is true or not, we have to get clear about what sort of creature it is. According to Little and Lance, ceteris paribus statements have the following two general characteristics (61-2):

(**A**) Non-empty claims about **the nature of things**: Although the ceteris paribus clause cannot be spelled out fully, ceteris paribus statements don't seem empty. They don't seem to be saying that they are true unless false. Rather, the ceteris paribus statements reveal something about the *nature* of things. For instance, the claim "ceteris paribus, fish eggs turn into fish" reveals, according to Little and Lance, something about the nature of fish eggs. Similarly, the claim "ceteris paribus, lying is wrong" reveals something about the nature of lying.

(**B**) Non-statistical claims about **privileged conditions**: Ceteris paribus claims are not statistical claims. They are not saying that in most cases, fish eggs turn into fish. Even if most fish eggs don't develop into fish, it is still true that ceteris paribus, fish eggs turn into fish. Rather, ceteris paribus claims are the so-called *lawlike* claims, sustaining a counterfactual statement between fish eggs and fish---for all x, if x were a fish egg, x would turn into fish. Little and Lance construes the lawlike claims in terms of what they call the "privileged conditions". To argue that "fish eggs turn into fish" expresses a lawlike connection between the two is to argue that in privileged conditions, fish eggs turn into fish. Similarly, it may well be contended that there is a lawlike connection between lying and the property of wrongness. In privileged conditions, lying is wrong. And the claim "ceteris paribus, lying is wrong" can be so construed.

In what follows, I have two respective points to make about (A) and (B). First, I will argue, contra (A), that ceteris paribus moral statements do not reveal the nature of a feature and that whether they are true depends on the context in which the specified feature is embedded in. Second, I will argue, contra (B) that there is no compelling reason for us to believe that ceteris paribus moral statements, when construed in terms of privileged conditions, are true. For the notion "privileged conditions" can be unpacked in such a way that in them lying is not wrong.

II. CRITIQUE OF (A)

Is it in the nature of the feature of lying that it has the property of wrongness-making? As we can tell from (A), Little and Lance give a "yes" answer by drawing an analogy to a biological case of fish eggs turning into fish. However, I believe that a dis-analogy exists between the biological case and the moral case. It is generally agreed it is not merely accidental that fish eggs turn into fish. There must be something in the nature of fish eggs that enables them to do so---such as the genetic make-up of fish eggs. But does the feature of lying have an intrinsic nature such that it always has the property of wrongness-making without exceptions? I think the answer is "no" due to the embedded character of the feature of lying:

Embedded Character: The feature of lying is always embedded in a context which involves at least the following factors: a liar, her motive for lying, a person being lied to and the consequences of the lie.

All of the above-mentioned factors can have an impact on the moral status of the feature of lying in some suitable circumstances. Whether the feature of lying has the property of wrongness-making depends on the context it is embedded in. If it is embedded in a context of a Taiwanese card game called "Bluffing", where telling a lie is part of the game, lying is not only not wrong but also not wrong-making at all. If "the nature of the feature of lying" conceptually requires that the feature of lying has to be always wrong-making, then the game "Bluffing" shows that the feature of lying cannot have such a nature.

It may well be contended here that only when a feature of lying is abstracted away from its embedded context, its true nature can be revealed. For only then can we be certain that the feature of lying is examined in its own light without the interference of other factors. Following these lines of thought, it may well be contended that when Little and Lance argue that lying is always wrong-making, they should be construed as endorsing

The Thesis of Abstraction: The feature of lying, when abstracted away from its embedded context, is always wrong-making.

Construing Little and Lance as endorsing the above thesis can line up with their view that it is in the nature of fish eggs that they turn into fish. For it is certainly not the case that fish eggs turn into fish in any context. Fish eggs might end up being in the bellies of big fish or they may well be destroyed by poisonous pollutants in the river. There are various sorts of contexts where fish eggs do not turn into fish due to the presence of some interfering factors. However, the claim that it is in the nature of fish eggs that they turn into fish still seems true. Why? I think this is because the claim has to be construed as holding in an "abstracting context", in which fish eggs are abstracted away from the interfering factors. In a biological laboratory, we can create an "abstracting context" such that fish eggs can be abstracted away from predators and poisonous pollutants, etc. In this sort of context where the interfering factors are absent, fish eggs turn into fish. Can an "abstracting context" also be created in the moral case? I think the answer is "no". There is no "moral laboratory" in which we can design a context such that the feature of lying can be abstracted away from the interfering factors. This is the place where the dis-analogy sets in between the moral case and the biological one. Fish eggs are still fish eggs when they are abstracted from their embedded rivers. They will turn into fish in well-controlled biological laboratories. By contrast, it is hard to imagine what a feature of lying would be like when it is shredded of its embedded context involving factors such as a liar, her motive for lying, a person being lied to, and the consequences of the lie. Namely, it is not clear such an abstracted feature of lying exists. Even if it does, it is not clear what it is that determines its moral status, given that it is now context-independent, not involving the above-mentioned factors. The claim that the feature of lying is always wrong-making is thus not warranted.

To sum up, given the dis-analogy between fish eggs and the feature of lying, I think that Little and Lance's argument by analogy fails. Although it might well be in the nature of fish eggs that they turn into fish, it is not so clear that it is in the nature of lying that it always has the property of wrongness-making.

III. CRITIQUE OF (B)

In order to assess whether it is true that in privileged conditions, lying is wrong, we need to first of all have a good grasp of what "privileged conditions" mean. According to Lance and Little, it can mean the following two things (p. 62):

(1) conditions in which an item's nature is revealed.

(2) conditions in which an item has the specified property.

When construed in terms of (1), "in privileged conditions, lying is wrong" should be understood as saying "in conditions in which the nature of lying is revealed, lying is wrong". However, when thus construed, the statement suggests that there might be conditions in which the nature of lying is not revealed, the *un*privileged conditions. In those unprivileged conditions, lying may not be wrong. In the Introduction, we have seen that Little and Lance argue that lying may not be wrong in all cases; however, even in those cases where it is not wrong, they think that it is still in the nature of lying that it is wrong-making. Here, we can rehearse our critique of (A) and contend that lying has no such a nature such that it is always wrong-making independent of the contexts it is embedded in. Hence, when the "privileged" claim is cashed out in terms of (1), its claim to truth is not warranted.

When construed in terms of (2), "in privileged conditions, lying is wrong" should be understood as saying "in conditions in which lying has the property of wrongness, lying is wrong." As we can tell immediately, this way of understanding "privileged conditions" will only make the "privileged" claim trivially true and rather uninformative. For we want to know in what sort of conditions, lying is wrong. So unless there is an independent way of telling the conditions in which an item has the specified property from those in which it does not, cashing out the "privileged" claim in terms of (2) is rather un-illuminating.

Little and Lance do think that there is an independent way of telling whether a condition is one in which an item has the specified property. They appeal to some practical understanding of a concept to distinguish those conditions in which an item has the specified property from the others (62). Namely, they endorse what I call

The Thesis of Practical Understanding: A good practical understanding of a concept involves a good practical understanding of the conditions in which an item has the specified property and those in which it does not.

The above thesis needs some explication. Let me illustrate by using an example provided by Little and Lance themselves. They think that a good practical understanding of the concept *match* involves a good practical understanding of the conditions in which it will light when struck and those in which it won't. However, they do not tell us which of these two sorts of conditions are the privileged ones. In fact, they claim that

> "[A concept of] the artifact kind, where deployed, is circumscribed by marking some conditions---however frequent or rare they may contingently be---as privileged." (62)

According to the quoted passage, one can, if one will, mark the conditions in which matches don't light when struck as privileged, even if these conditions are rare. The fact that we don't mark them this way does not change the fact that we can do so. Similarly, one

might as well mark those conditions in which lying is right as privileged, even when such conditions are rare. When the privileged conditions of lying is so marked, then the claim "in privileged conditions, lying is wrong" is apparently falsified.

It might well be objected that when the privileged conditions are so marked, they fail to square with the following empirical fact: those who have a good grasp of the privileged conditions "generally succeed in their attempts to light matches, don't waste their matches by making attempts when there is no hope of success" (p. 63). However, I think the objection is toothless due to

The Quinean Thesis of Underdetermination: The same empirical data might well be explained by various different and incompatible hypotheses.

To illustrate with an example Little and Lance used, they argue that the empirical data that denizens of watery Atlantis generally succeed in their attempts to light the matches when they relocate themselves to the Australian outback is explained by their grasp of the concept *match* and their marking as privileged those conditions in which matches light when struck. However, I think that the empirical data may well be explained by the hypothesis that being in the Australian outback they know that they are now in what they regard as *un*privileged circumstances where matches light when struck. In other words, to explain their sensible practices, it is not necessary to mark as privileged those conditions in which matches light when struck. It will also do to mark those conditions as *un*privileged and those conditions in which matches don't light when struck as privileged. When the distinction between privileged and unprivileged conditions is thus made, their concept *match* can still be functional. When they relocate themselves again back to the watery Atlantis, which they regard as a *privileged* environment where matches don't light when struck, they don't waste their time striking matches as they clearly know "there is no hope of success".

IV. CONCLUSION

In this paper, I have demonstrated that Little and Lance did not provide strong arguments for us to believe that ceteris paribus moral principles, when understood as expressing non-empty claims about the nature of things or non-statistical claims about privileged conditions, are true. If so, is it really the case that any adequate morality has to recognize the truth of ceteris paribus moral principles, as Little and Lance suggested? Until more compelling arguments for the truth of ceteris paribus moral principles are produced, I think we are justified in remaining skeptical.[1]

tsu.shiu-hwa@anu.edu.au

1] I am grateful to Jeanette Kennett for her comments on the earlier drafts of this paper and to the audience at the 2009 postgraduate seminar at the Centre for Applied Philosophy and Public Ethics, Australian National University for relevant discussions.

REFERENCES

Little, Margaret and Mark Lance. 2006. Defending Moral Particularism. In *Contemporary Debates in Moral Theory*, ed. James Dreier, 305-21. Blackwell, Oxford.

———. 2008. From Particularism to Defeasibility in Ethics. In *Challenging Moral Particularism*, ed. Matjaž Potrc, Vojko Strahovnik and Mark Lance, 53-75. Routledge, New York.

Public Reason 1 (2): 94-10

Book Reviews

Cavarero, Adriana. 2009. Horrorism: Naming Contemporary Violence. Trans. William McCuaig. New York: Columbia University Press. Pp. 154. ISBN: 0231144563

How to name the constellation of violence, power and resistance that characterizes the contemporary political scene? Are the traditional political categories sufficient for a representation of our contemporaneity? Can the language of this tradition aptly describe and interpret what is happening today? These questions inform Adriana Cavarero's new book and lead her to attempt a renaming of the phenomenon of contemporary violence. Language, in fact, has proven unable to renew itself in order to represent, and thus comprehend, the global carnage that stains the beginning of the twenty-first century; indeed, she writes, "it tends to mask it" (2). In the twentieth century violence spread and assumed unheard-of forms, and since September 11, 2001, it marks the global everyday life in a way that escapes the old interpretive frameworks. We have no words to describe a form of violence that strikes everywhere, at any time, and mainly defenceless civilians: the concepts from the past, like *war* or *terrorism* misleadingly confine this violence into categories unable to represent the new. Linguistic innovation becomes therefore imperative and Cavarero proposes to situate the new phenomenon in the semantic field of *horror*: the neologism "horrorism," apart from the obvious assonance with the word terrorism, is meant to emphasize "the peculiarly repugnant character of so many scenes of contemporary violence" (29). In an analysis that spans from Greek mythology, through the main political thinkers of modernity like Hobbes, Schmitt, Foucault and Arendt, the horrors of Auschwitz and Bataille's eroticization of violence, Primo Levi and Joseph Conrad, to suicide bombers and the tortures at Abu Grahib, Cavarero unravels the roots, iconography and poignant actuality of contemporary "horrorism."

This renaming entails primarily a change of perspective: from the traditional viewpoint – what Cavarero names the "perspective" or "criterion of the warrior" – the new form of violence remains incomprehensible and unrepresentable. It is the defenceless person without qualities, blown up by suicide bombers, bombed by unmanned aircrafts, tortured, raped, displaced, confined into camps, who takes the centre of the contemporary stage, and it is only from her perspective – the "criterion of the defenceless" – that the phenomenon must be named and described. It makes no longer sense, for example, to discuss war in terms of regulated conflicts between states and the classical model of a clash between men in uniform. From its first account in the Homeric battle, this model entail reciprocal, symmetrical violence, and not unilateral violence inflicted upon the defenceless. *Reciprocity* is its fundamental principle, and terror is its essence. From at least the Armenian genocide and World War I, war has become not only asymmetrical, but "consists predominantly of the homicide, unilateral and sometimes planned, of the defenceless" (62). When most of the 'casualties' of war are helpless civilians, it is impossible and senseless to ignore their point of view and still entrust the meaning of war and its horror to the perspective of the warrior. It makes no sense to insist on the criterion of the regularity of combatants, when the victims of any war are now civilians by a wide majority. Carl Schmitt attempted a redefinition of modern war and the concept of enemy in his *Theory of the Partisan*, but in order to 'update' the criterion of the warrior, and certainly not to reverse it. It is equally senseless, today, to

separate strategy and goals, means and ends: from the point of view of the helpless victim, "the end melts away, and the means becomes substance" (1). It is precisely this distinction that opens the book: two scenes of massacre, a suicide attack in Baghdad in July 2005 and the American bombing of a wedding feast in Iraq in May 2004, a 'mistake,' are inserted, from the perspective of the warrior, in a narrative that finally justifies the massacre either as part of a strategy to achieve 'higher ends' (as deplorable as they may be) or as 'collateral damage,' deplored but inevitable facts of war. It is only from the perspective of the helpless victims that this narrative can be shown not only to be hollow and ambiguous, but to provide the linguistic justification of what can only be described as "crime" (3). It is the horror of the scene that stands out, and from this horror a new conceptual and political framework must arise.

In the age of the 'war on terror,' the distinction between war and terrorism is a crucial problem: within the traditional framework, terrorism is defined as a criminal form of violence, whose actors, aims and acts are incompatible with the traditional system of destruction. The terrorist is no regular combatant who directs its fire against other combatants, hitting civilians only by mistake: to kill civilian is today most often the goal. This framework functions of course in the discourse of politicians and the media as a legitimization of 'just,' 'preventive,' or even 'humanitarian' wars, and it is certainly not oblivious of the enormity and suffering of civilian victims. These, however, are given neither a place that accounts for their status nor a voice to represent it. Cavarero emphasizes then that, though the label 'terrorism' functions as an umbrella concept which groups a plethora of historical phenomena, and though all these phenomena are characterized by the massacre of the defenceless, it is only in today's development that the weapon becomes the body of a suicide. The *terror* becomes thus *horror*. The terror which characterizes contemporary violence has lost its goals and thus cannot be defined as strategic. The fact that the weapon becomes the body itself is not only scandalous, but, from the point of view of today's technological imaginary of war, irregular, illegitimate and also unfair. Not only there is no longer symmetry between combatants, but there isn't even any battle. The omnipotent dreams of military hypertechnology and the very concept of war the regular combatants still maintain they are fighting, shine for their emptiness. The enemy itself has become an indistinct, phantom-like shadow, indistinguishable and unrepresentable. And torture, as epitomized by the pictures of Abu Grahib, reveals the mere horrorist face of a violence devoid, in both camps, of any goal or strategy. Finally, the figure of the victim has grown global: victim can be anyone at all, an indiscriminate and random 'casualty.' The old framework is, therefore, not only extremely ambiguous, but its argumentation never goes so far as to embrace radically the criterion of the defenceless.

Horror is not, of course, a novelty in the universal history of violence, and Cavarero goes a long way to retrace its semantic and iconographic roots in Greek mythology – a trademark of her writing. This etymological operation responds to a two-fold strategy: firstly, it allows Cavarero to make a clear distinction between 'horror' and 'terror,' between their characters and manifestations and their effects on the body. Whereas 'terror' derives from the Greek and Latin verb *tremo* and connotes a fear that "acts immediately on the body, making it tremble and compelling it to take flight" (4), 'horror' comes from *horreo* and denotes primarily a state of paralysis, which excludes the moment of flight. Violent death is part of the picture, but not the central part: "There is no question of evading death. In contrast to what occurs with terror, in horror there is no instinctive

movement of flight in order to survive" (8). Moreover, if 'terror' is unequivocally related to fear and fright, 'horror' has more to do with repugnance, a repugnance that is mainly related to the sight of a dismembered body. Horror denotes a scene unbearable to look at, like that of bodies that blow themselves up in order to tear other bodies apart, dismembering their own individuality and that of their victims. Contemporary violence – and this is the second point – has taken such a form that mainly attacks the integrity of the human body – suicide bombers, beheadings, mutilations, torture – and thus escapes the traditional vocabulary of war based rather on the semantics and 'physics' of terror. It is the terminological constellation of horror, Cavarero argues, that we need to use in order to describe and comprehend this new form of violence.

The excursus into Greek mythology allows her to make another point: in the iconography of the misogynist, patriarchal West, it is two feminine figures, Medusa and Medea, which epitomize horror. Medusa, the Gorgon whose gaze could petrify and was finally beheaded by Perseus, and Medea, wife of Jason, who first killed and dismembered her brother and then killed her two children in revenge for Jason's betrayal: "Horror has the face of a woman" (14). The severed head of Medusa symbolizes not only the unwatchable dismemberment of the body, but also the horror of the separation of the female head from the uterus and its reproductive function, to which the patriarchal narrative relegates women. Medea, killing her children, emphasizes not only the horror of a violence inflicted to the helpless *par excellence*, but also the horror of a woman that renounces her stereotypical reproductive function and gives death instead of life. If men remain unchallenged protagonists in every theatre of violence, "when a woman steps onto the stage the scene turns darker" (14). The horror of contemporary female 'terrorists' and female suicide bombers – some even pregnant – evokes these two figures of the patriarchal iconography, but simultaneously also disarranges the gender dynamics of the traditional (male) imaginary of war and violence, and emphasizes once again the insufficiency of its categorical framework.

The fact that it is the very singularity of the victim that becomes accidental spells out the fundamental issue that the criterion of the helpless identifies: the superfluity of the human being. Horrorist violence, by tearing furiously at the body, works not simply to take away its life, but to "undo its figural unity" and thus emphasizes that it is the uniqueness of the person that is being attacked (15). In other words, this is a violence that goes beyond death and whose goal is not much death but the destruction of human singularity in its ontological dignity. Its figure is in fact the severed head of Medusa, epitome of a body dismembered, undone and disfigured, and thus attacked in its irremediable incarnated singularity. Most repugnant than any other body part is the severed head, the most markedly human of the remains, on which the singular face can still be seen: "Medusa alludes to a human essence that, deformed in its very being, contemplates the unprecedented act of its own dehumanization. The quintessence of an incarnated uniqueness that, in expressing itself, exposes itself, the severed head is the symbol of that which extreme violence has chosen for its object" (16). This body is unwatchable and arises instinctive disgust for a violence that, not content merely to kill because killing would be too little, aims to destroy the uniqueness of the body: "What is at stake is not the end of a human life but the human condition itself, as incarnated in the singularity of vulnerable bodies" (8). A clear example is modern beheading: the crime is staged as an intentional offense to the ontological dignity of the victim. The question is not so much killing but rather "dehumanizing and savaging the body as

body, destroying it in its figural unity, sullying it" (9). And this extreme violence, directed at nullifying human beings even more than at killing them, relies on the semantics of horror rather than that of terror. What this violence really perpetrates, therefore, is an "ontological crime," one whose "precise aim is to erase singularity" (19), one whose goal is the killing of uniqueness.

The slaughter of the defenceless is not a specialty of modernity, but the history of the twentieth century stages the ontological crime in forms and proportions never achieved before. Beginning with the genocide of the Armenians in 1915-1916 and the unheard-of carnage of World War I, the "short century" takes the killing of uniqueness to organisational and technological perfection. The apex – though sure enough not the last instance – of horrorism was reached with the Nazi death camps. Auschwitz epitomizes this horror insofar as it construed a system for the fabrication of the degenerated helpless person and thus constitutes an "exercise of demolition of the human being" (36). Cavarero reads Primo Levi's poignant pages on his experience in the camp through the theoretical lens of Hannah Arendt's *The Origins of Totalitarianism* in order to emphasize that what the camp really epitomizes is an attack on the ontological status of the human being: it systematically aims at transforming unique beings into a mass of superfluous, impersonal beings whose murder is so impersonal that "also takes away from them their own death" (43). This system is finally paradoxical because its end-product is the *Muselmann*, the human reduced to 'bare life,' no longer exposed to offence because by now incapable of suffering, and thus no longer vulnerable: "they can no longer even feel the hurt of the *vulnus* that nevertheless continues to be inflicted on them with methodical perseverance" (34). The *Muselmann*, the outmost figure of almost grotesque helplessness, is paradoxically *invulnerable*, she signifies a stage of so extreme defencelessness that even vulnerability has been taken away from it. The emphasis is here, however, not on the question of *zoé*, *bios* or 'bare life,' but rather on the ontological dimension and significance of the system: "Extreme horror […] has to do with the human condition as such" (43). The violence of the Lager is essentially aimed at "fabricating a victim, insensitive by now to the *vulnus*, in whom the human dignity of the defenceless degenerates into a caricature of itself" (36). The issue is therefore, Cavarero insists, not only ethical or political, but involves first and foremost the question of ontology: it is human nature as singular, unique and incarnated body, that is concerned.

This is a concern that Cavarero, with Arendt, carries to a wider philosophical level. The attack on singularity as the ontological dignity of the human being is in fact, according to Arendt, what characterizes the history of Western philosophy, which sacrificed human plurality on the altar of the absolutisation of the One. Ignoring men in flesh and blood, and thus erasing their uniqueness, particularity and finitude, the philosophical tradition fabricated a series of abstract 'fictitious entities' which finally made the concrete human being 'superfluous.' And the idea of the superfluity of the singular is what informs the horror of so many forms of politics. Nazism, in sum, put into operation what philosophy had only thought, and 'fabricated' the superfluity of human beings. This notion of the superfluity of the individual also informs Georges Bataille's eroticisation of violence, and since many suggestions arising from his work still burden the contemporary understanding of violence – Cavarero cites as an example James Hillman's 2004 book *A Terrible Love of War* – Bataille's arguments are thoroughly dissected in the book. The dissolution of the finite into the infinite, erotically enjoyed in cruelty, is the focus of his literary and theoretical constructions: his "sovereign subject" is he who, in

contrast to the servile subject, does not follow the bourgeois principles of utility and self-preservation, but rather those of loss and self-destruction, experiencing full erotic inebriation: "against the instinct of self-preservation seen as an act wherein the I closes in on itself, it is the death wish that defines the liberty of the sovereign soul" (51). The gallery of this enthusiastic dissolution of the 'I' is the horror house of Bataille's imaginary: bodies raped, flayed, dismembered, whose disfigurement ruptures the boundaries and nullifies the singularity of the human being. There is, significantly, no *reciprocity* in this relation to the other, but most of all what this erotic deindividualisation shuts off is the vision of the fundamental alternative that vulnerability offers, that between wound and care.

The criterion of the helpless, in fact, not only provides the theoretical instruments to describe and represent contemporary violence, but also functions as ethical and political standpoint. A trademark of Cavarero's thought is her relying on an ontology of uniqueness and exposure that she derives from Arendt and then develops and combines with the feminist reflection known as *pensiero della differenza sessuale* (theory of sexual difference). In *Horrorism*, this ontology is developed along the lines of Judith Butler's reflections on "vulnerability" in *Precarious Life*. Vulnerability is one of the constitutive characters of a unique being exposed to the other: "The uniqueness that characterizes the ontological status of humans is also […] a constitutive vulnerability, especially when understood in corporeal terms" (20). To be unique means to be exposed to the other and to consign one's singularity to this exposure. The human, unique being is vulnerable by definition. The condition of vulnerability presents an essential alternative which moves between the two poles of wounding and caring: "Inasmuch as vulnerable, exposed to the other, the singular body is irremediably open to both responses" (20). For Butler, Cavarero emphasizes, vulnerability configures a human condition in which it is the relation to the other that counts and puts to the fore an ontology of linkage and dependence. Recognizing our common condition of vulnerability calls for a collective responsibility. This move entails a rejection of the autonomous sovereign subject of the Western philosophical and political tradition, which, like the sovereign state to which it corresponds, thinks of itself as closed and self-sufficient: against the individualistic modern ontology, which refuses to admit dependency and relationship, Butler emphasizes that the 'I' is not closed but rather open and exposed. And this exposure consigns primarily the subject to the *vulnus*, to the alternative between the wound that the other can inflict and the care that the other can provide. The vulnerable being "exists totally in the tension generated by this alternative" (30).

Cavarero points out, however, that 'vulnerability' is not a synonymous of 'helplessness.' The human being is vulnerable as a singular body exposed to the wounding. Yet, there is nothing necessary in this vulnerability, only the contingent potential for the wound. "As a body, the vulnerable one remains vulnerable as long as she lives, exposed at any instant to the *vulnus*" (30). 'Helpless' presents a different and stronger connotation: the Italian word employed by Cavarero, and that the English translator chose to render indifferently as "helpless" or "defenceless," is *inerme*, which etymologically means 'unarmed,' he who does not bear arms and thus cannot harm, kill or wound. In everyday use the term tends to designate a person who, attacked, has no arms with which to defend themselves. To be defenceless means to be in the power of the other and thus entails a condition of substantial passivity. The relation is unilateral, there is no reciprocity, no symmetry, no parity. The exemplary case is the infant: the defence-

lessness of a baby does not depend on circumstances, but is a condition, the essential mode in which the human being comes into the world and, for a certain period, inhabits it. Infancy is the span of time in which vulnerability and helplessness are completely conjoined: "Though she remains vulnerable as long as she lives, from the first to the last day of her singular existence, an adult falls back into defencelessness only in certain circumstances. She is always vulnerable but only sometimes helpless, as contingency dictates and with a variable degree of intensity" (30-1). In the infant, the relation takes the form of unilateral exposure: "The vulnerable being is here the absolutely exposed and helpless one who is awaiting care and has no means to defend itself against wounding. Its relation to the other is a total consignment of its corporeal singularity in a context that does not allow for reciprocity" (21). It is precisely the thematisation of infancy that allows the vulnerable being to be read in terms of a drastic alternative between violence and care: the other, embodied here by the mother, cannot limit the care to a mere refraining from wounding, but, by necessity, "the vulnerability of the infant always summons her active involvement" (24). The infant thus proclaims relationship as a human condition not just fundamental, but structurally necessary.

The gloomy landscape of the twentieth and twenty-first century has transformed the contingency of helplessness into necessity: the circumstances that produce helplessness have dilated into the indeterminacy of a space and a time corresponding to "the everyday dimension of the everywhere" (75). More than circumstances, we can speak of an ongoing condition which makes vulnerability coincide with helplessness: "Exposed unilaterally to the *vulnus*, the defenceless are the targets of a violent death that surpasses the event, atrocious in itself, of death, because it has degraded each of them beforehand from singular being to random being" (76). Therefore, the viewpoint of the defenceless, Cavarero argues, must be adopted *exclusively*: not merely as the only prospective from which contemporary violence can be really named, represented and understood, but also that from which subjectivity, relationality, ethics and politics must be rethought.

Carlo Salzani
Rheinische Friedrich-Wilhelms-Universität Bonn

Müller, Jan-Werner. 2007. Constitutional Patriotism. Princeton University Press. Pp. 186. ISBN: 978-1-4008-2808-1

The concept of constitutional patriotism is not Jürgen Habermas's, even if it has come to be associated with his version of the post-national state. The term itself, as Jan-Werner Müller points out in this important work, was not coined by Habermas, but by Dolf Sternberger, a student of Hannah Arendt, to describe the ideal relationship between the German state and its citizens in the 1970s.

The idea, as Müller traces its history, begins with Karl Jaspers's *The Question of German Guilt*. While Jaspers rejected the idea that the German people were collectively guilty, he believed nonetheless that they were in some way collectively responsible for the Holocaust. This was not necessarily a negative outcome: if the German people shouldered that responsibility – a responsibility for the worst criminal act in history – "a negative past could become a source of social cohesion" (16). While constitutional patriotism shares similar characteristics to other methods of achieving social cohesion, such as a shared national narrative (characteristics such as a concern with memory and militancy), it differs from them by emphasizing a different social imaginary (in this

case, a repudiated past – a history that could have been otherwise, but serves as a source of instruction for the current generation).

However, while the idea of constitutional patriotism may have emerged in a specifically German context – as a way of addressing Germany's exceptional 20th century history – Müller believes that it is applicable beyond these narrow historical particulars.

What makes Müller's work so important is that it is the first book I know of to try to develop a theory of constitutional patriotism that, while clearly informed by Habermas's writings, attempts to develop an independent justification for the idea. He is ideally suited to the task: of contemporary political theorists, Müller has perhaps devoted the most effort to this project; chapters of the book have been adapted from articles published in the journals *Constellations* and *Contemporary Political Theory*, amongst others. Müller's goal in the book is to show that constitutional patriotism offers a middle ground between cosmopolitanism on the one hand, and liberal nationalism on the other – concern for all human beings everywhere, versus concern for one's co-nationals only. Müller tries to show that constitutional patriotism captures the best of the need to motivate political agency by reference to particular experiences and concerns, without giving up a set of universalist norms. In this review I treat Müller's efforts to develop such a theory; I leave to others a discussion of his application of that theory, in the final chapter, to the European Union.

First, Müller sets down the rules of engagement: constitutional patriotism is not a theory of the self, nor a theory of justice, but a way of maintaining the liberal state. Müller concedes much to his opponents when he announces that first, "constitutional patriotism is...not by itself some kind of civic panacea in cases of collective political breakdown," and that second, it "cannot by itself generate large degrees of social solidarity" (48).

It is widely conceded in political theory that some sense of belonging to the same historical community is a prerequisite for obtaining social rights in a liberal democracy; these rights cannot simply be the result of the application of some rule. Additionally, the rule of law can extend only over some defined territory. Finally, in a liberal democracy, the laws of the state must generate some normative sway beyond mere coercive force.

The classical solution to the legitimation problem has often been to advocate some sort of nationalism. In some cases, this is out of necessity a form of patriotism. Recent thinkers, such as George Kateb, have argued that patriotism is out of necessity an illiberal form of group meaningfulness. Against this, the proponents of constitutional patriotism (including Müller and Karol Edward Sołtan) argue that this is to pigeonhole patriotism: constitutional patriotism, in contradistinction to Kateb's understanding of patriotism, is both a form of commitment to the universal principles of modern constitutions and human rights, and to one's own state. Thus, constitutional patriotism is not a loyalty limited to one particular state or nation (the possibility of some form of cosmopolitan constitutional patriotism will always remain open).

The important difference, as Müller tries to show, is that unlike traditional patriotism, which is loyalty to a nation or national history, constitutional patriotism is a loyalty to a way of living in a community. It thus generates social cohesion, he argues, without the problems of national chauvinism.

After discussing the historical background of the problem and motivating the theory of constitutional patriotism in the first chapter, Müller turns his attention to respond to various critics of his project. Having addressed the twin criticisms that con-

stitutional patriotism "is too abstract" or "not enough blood in it for me" (49) in his defence of constitutional patriotism against liberal nationalism, he wants to show in the second chapter that constitutional patriotism is neither a form of statist nationalism or a kind of civic religion, either of which could presumably be ethically dangerous.

Against the charge that it is a form of statist nationalism, Müller begins by arguing that an attachment to the idea of a constitution (not of any particular constitution) is a necessary condition for living in a shared society (in this way, he proceeds with a method of rational reconstruction not unlike Habermas's attempt to reconstruct the necessary conditions for political discourse). Only in this way, Müller argues, can the idea of the rule of law, and adherence to majority decisions when one is in the minority, seem palatable to members of the modern polity. Constitutions function to "produce a form of contained disagreement or limited diversity" of opinion (55). From debates over the form of the constitution and the state emerge a shared constitutional culture that serves as a glue to hold members of a society together. Unlike an identity, the self-understanding of this culture is framed against ever changing historical experiences, new information etc.

Thus, against this first charge, Müller argues that constitutional patriotism is not a question of identity, but of a shared commitment to work together to establish a stable body politic that respects the need to treat every other member as an equal – in other words, to identify all other members of a society as co-nationals. Rather than making membership in a particular community a condition for civic membership, constitutional patriotism demands that those wishing to be members of a liberal democracy view others as members of their own political communities. Constitutional patriots are not committed to a thing (the state), but to a process of living together.

Against the second charge, Müller wants to show that constitutional patriotism does not lead to civic religion. He begins by distinguishing three ways in which we might talk about civic religions. First, there is the idea that there might be some dominant religion that structures a modern society. Second, we might speak of a civic religion à la Rousseau: treat religion instrumentally as a means of integrating society. Thirdly, we might speak of a soft civic religion wherein the state sponsors certain types of historical commemorations to ensure civic pride: symbols of national pride such as flags, national anthems, pledges of allegiance, ceremonies at statesmen's tombs, etc. (81). Müller assumes, somewhat uncritically, that it is only the third understanding of civic religion that would concern (presumably cosmopolitan) critics of constitutional patriotism. These critics, on his telling, worry that the veneration of historical events, figures and memories, might encourage some form of uncritical citizenship. In particular, patterns of "veneration might encourage the strategic manipulation of constitutional symbols by political elites" (82).

Clearly, such constitution veneration is incompatible with a general theoretical understanding of constitutional patriotism. This, however, is too easy a way out. Müller argues that this is not the only, nor best, possible response: constitutional patriotism carries with it the normative resources to challenge such a blind veneration of symbols by insisting, rather than on identity, on a political culture that venerates process over substance.

I have two objections to this otherwise fine work; both rendered all the more vexing because Müller seems to have addressed them in other places. In a version of the

second chapter published in *Constellations* as "Three Objections to Constitutional Patriotism" (2007. *Constellations* (14): 195-206). Müller wonders whether or not constitutional patriotism is a form of modernism in some undesirable sort of way. Much of that discussion is left unelaborated on in the chapter. Put simply, Müller asks, following the objections of thinkers like Thomas Meyer, if "constitutional patriotism designates a particularly modern identity" (2007, 203). The obvious contrast would be both to pre-modern national and cultural identities, and with post-modern identities. The first, obviously, are inherently nationalistic in some strong way, and in that respect constitutional patriotism is clearly a modern approach.

However, the post-modern concern argues that other societies exist that take different polities (not necessarily constitutional democracy) as a starting point. Thus, constitutional patriotism is necessarily biased towards liberal democracies — why should not other post-modern identities be seen as early stages of genuine cosmopolitan citizenship? While I agree with Müller's contention in the article that this is to confuse constitutional patriotism with the embrace of actually existing constitutions (as opposed to its contingent nature), highlighting this objection in the book could only have served to underline what I think is Müller's important insight.

Second, following the idea developed in my first objection, had Müller stressed the important distinction between patriotism for the constitution (as a general idea) versus patriotism for one specific constitution, he would have been able to develop an important and overlooked parallel between Habermas's work and his own. In the same way as Habermas views validity as a condition of process and not result in his discourse theory, constitutional patriotism is the acceptance of a political process over any concrete history. It is this distinction that in my mind renders constitutional patriotism preferable to liberal nationalism in its various forms.

Part of the problem with constitutional patriotism has always been to try to show, in a normative account, what comes first: attachment to universalist values, which are then realized in some particular setting, or some particular polity (that is democratic in some essential way) that then be made an object of civic loyalty. In another place, Muller has written: "given the apparent tension between universalism and particularist loyalty, it is no wonder that critics have concluded that constitutional patriotism is simply an 'inconsistent idea' or just a kind of aspirational oxymoron, a well-meaning normative muddle, rather than a coherent normative proposal to rethink political solidarity and attachment" (2006. "On the Origins of Constitutional Patriotism". *Contemporary Political Theory* 3 (5): 278-96; 73). In other words, is there some sort of underlying normative framework which could give rise to a theory of constitutional patriotism or is it just a reconstruction of already underway moral developments in twenty-first-century societies.

While Habermas would see no disjunction here, Müller the aforementioned article is not content to continue in the tradition of critical theory. He wants to show that out of a particular German situation can arise a more robust theory. His constitutional patriotism wants to show that two things are possible: attachment to universal norms and a constitutional culture.

Müller in his book however does not take this route. He is content, as I have argued above, to proceed by way of rational reconstruction. As he does not address his (slightly) earlier paper, the reader has no way of knowing if he has revised his earlier

Rawlsian attempts to construct a freestanding justification of constitutional patriotism or not.

None of this, however, is to take away from what is undoubtedly a fine and important contribution to political theory.

Kevin William Gray
American University of Sharjah

Dhanda, Meena, ed. 2008. Reservations for Women, India: Issues in Contemporary Indian Feminism, v. 6. New Delhi: Women Unlimited. Pp. 390. ISBN 81-88965-41-3.

Reservations for Women comes as a precious source among the main writings dedicated to the political position of women in India. It is the first collection of essays and writings that addresses the issue of affirmative action as a way of increasing the presence of women in the Indian legislative assemblies. *Reservations for Women* belongs to the "Issues in Contemporary Indian Feminism" series, whose aim is to facilitate access for scholars, teachers and activists to all important writings related to gender.

Although it could be regarded as a natural symptom of the wider global context of female political locations, *Reservations for Women* also emphasizes the particularities of the Indian feminist movement.

Feminism in India is to be understood differently from Western feminism for many reasons. Firstly, the Indian woman is an epitome of the Indian culture; therefore, a feminist stance would in fact be equal to a nationalist stance. Secondly, although mainly patriarchal society (with some exceptions), Indian society has reserved a special place for women within culture through religious figures. Lastly, Indian women have defined themselves in harmony with the collective, not in opposition; indeed, in a collectively-oriented society such as India, feminism cannot be defined by individualism.

Another distinguishing element is that one cannot really speak about female oppression by men in India. On the one hand, this is because the Indian religion renders women complementary and equal to man. On the other hand, it was in fact men who initiated several social movements to improve the conditions of women in India (e.g. the abolition of the practice of *sati*). Lastly, one could argue the hierarchies among women are even stricter and more oppressive as a result of caste relations.

In spite of the above-mentioned challenges, Meena Dhanda, head of Philosophy at the University of Wolverhampton (where she has taught since 1992), aims to select the most important writings in the field of Indian political representation of women, without claiming to bring new arguments into light: "Much remains to be done and said; this book brings together what has already been said." (xvii) In a society where the caste system and communalism are major features, multiple patriarchies lead to multiple feminisms. This is why Dhanda insists that, in a heterogeneous theoretical environment, a selection of the major writings would impose a common framework for discussion.

In the introductory essay, Dhanda explains the place of this book in the Indian tradition: "Political thinkers now agree on the need for greater political participation of women. The disagreements now, as then, are about how to bring about the desired change." (xiv) In short, the volume presents different positions on the Women's Reservation Bill and suggestions regarding new methods for female political participation.

The book has four main sections. The first section presents the main divergent

views formulated by leaders of pre-independent India (1930-40's) and after. The focus of the discussion is on the divide of caste. The section begins with the *Declaration* of Naidu and Nawaz, addressed to the British Prime Minister in 1931 – where the two authors express a refusal of preferential treatment within the Parliament. The following contribution is a note of dissent to the *Report of Women in India*, written 40 years after, by Sarkar and Mazumdar (1974). The two authors insist that the reservation of seats for women must extend to legislative bodies too, and not be applicable to local administration only. A later comment by Mazumdar follows. The historical section ends with Mary E John's depiction of the last century; John emphasizes the tension between the reservation based on caste and the reservation based on gender.

The second section focuses on theoretical issues. This collection of theoretical writings addresses the legitimacy of representing women as a group, issues in defining inclusiveness in democracy, and puts into question the necessity and efficiency of top-down measures. The section begins with an excerpt from *The Quota Question* (Gandhi and Shah) and continues with one of the most cited works on the subject of gender quotas – Anne Philips' chapter from *The Politics of Presence*. Rai challenges the legitimacy of any quota policy, but not the idea of representation in itself; furthermore, she recommends the language of empowerment as instrument for female representation. The use of the language of empowerment is later criticized by Menon, since its meanings are too narrow to fit the Indian context. Last but not least, Dhanda suggests that the focus must be shifted from analyzing the consequences of policy of gender quotas to "what does it mean to engender democratic participation" (132).

The third section, called "Women as Policy Makers", is a case study of women's contributions to the 73rd and 74th Constitutional Amendments, and to female political representation in general. Omvedt fights the preconception that women would be only puppets in men's hands when politically elected, citing the experience of women in *panchayats* (local assemblies) in Maharahtra. Lama-Rewal reviews a survey conducted in 2000, which aims to show the concrete results of the Women's Reservation Bill. She concludes that "more and more women contest – and win – against male opponents." (xxvi); nevertheless, an increased female participation does not lead to a change in the House's agenda. This final point is rejected by Chattopadhyay and Duflo, whose research concludes that female representation ensures adequate delivery of public goods to disadvantaged categories. The section ends with several examples of good practice provided by Geetha.

The last section presents several alternatives to the Women Representation Bill (WRB). Kishwar's most cited work, "Women and Politics beyond Quotas", is here reproduced, along with Raman's commentary. Raman believes that ways more subtle than WRB must be found, because the WRB eventually strengthens the interests of the dominant groups. Narayan et al. present the most detailed alternative to the WRB – the Alternative Bill – while identifying its flaws (e.g. the rotation of constituency). By contrast, Omvedt praises the Alternative Bill, but suggests that a system of proportional representation (PR) is more appropriate. Lastly, Nanivadekar argues for the implementation of a dual-member constituency.

Dhanda's conclusion is that while theoretical and political debate must be continued in order to find the best solution, women representation must be implemented, even if through the (less-refined) Women Representation Bill. Apart from this, *Reservations for Women* does not focus on furthering the discussion. But in synthesizing the

main historical arguments, it serves as a precious resource for contemporary proponents of women's enhanced position in the Indian political life.

Diana Constantinescu

Public Reason 1 (2): 106

Book Notes

Stan, Lavinia, ed. 2009. Transitional Justice in Eastern Europe and the Soviet Union. London: Routledge. Pp. 328. ISBN: 978-0-415-77671-4

For students of transitional justice, be they scholars, policy-makers or the general public, this recent book of Lavinia Stan (editor and chief author) is a timely guide in advancing our understanding of the complex phenomenon of transitional justice. In the field of studying the coming to terms with the dictatorial past in democratising countries, a field which has rapidly expanded over the last 30 years, this book fills important gaps. More than a simple collection of individual country studies, Transitional Justice… offers a coherent vision and significantly advances the knowledge in the field.

The book tackles courageously he fundamental question of why some countries choose to deal with past repression (by opening up the archives of the secret police, excluding the representatives of the previous regime from public office, and prosecuting human rights abuses through court trials), while other countries with similar histories of human rights abuses do little to face their past?

As its title indicates, the book deals with the European post-communist region. This region has been partly neglected by the transitional justice literature, which had tended to study in detail countries that have conducted important and transitional justice policies early on in the post-communist period (Germany, Czechoslovakia, but also Poland). Lavinia Stan, however, proves that "non-cases" are as relevant as the exemplary cases for understanding the dynamics of transitional justice. Among the most significant contributions of the book are its detailed studies of countries that were given little or no attention in transitional justice: cases like Albania, Slovenia and the former Soviet Union – the Baltic countries, but also Moldova, Georgia and a host of other former Soviet Republics whose official reckoning with the communist past is almost non-existent. Stan does also a fine presentation of the Romanian case – the best yet in the transitional justice literature. The best-known cases are also examined in great detail, following not only political intent or law promulgation, but also the actual implementation and concrete results of policies, from 1990 up until 2007 – a time span coverage unprecedented in the literature.

But what we believe to be the book's greatest achievement is its combination of meticulous attention to the particularities of each country and of a theoretical approach that truly offers a better understanding of the post-communist world and of the transitional justice phenomenon. As shown in a concluding section that would endure as a valuable advancement of the field, it is only from attentive observation of the peculiarities of each case that generalities can be drawn. The categories of the early "transitology" literature are shown to be, if not outright obsolete, at least lacking the finesse necessary to understand transitional justice policies and to predict outcomes. Not only the nature of the repression, or the way the regime ended, are relevant for transitional justice; to these factors must be added more complex and dynamic assessments of the "politics of the present," that is, the make-up of the political landscape at the beginning of transition, as well as its evolution along successive election cycles. Transitional justice was everywhere used as a weapon in the struggle for political power, and a strong connection can be established between the impetus of transitional justice measures and the political parties in government at that time. Lavinia Stan and the contributors give accurate accounts of this dimension and of several others, which combine in a multi-factorial model of explanation of the differences between apparently similar countries.

It is an important contribution, that will for many years be an essential reading for anyone who wants to know why coming to terms with the past is so difficult in the post-communist countries.

Raluca Ursachi
University of Paris Panthéon-Sorbonne

PUBLIC REASON

Journal of Political and Moral Philosophy

Submission information: All contributions whether articles or reviews are welcome and should be sent by e-mail as attachment at brancoveanu@publicreason.ro. Papers will be refereed on a blind basis by *Public Reason*'s referee board. Acceptance notices will be sent as soon as possible. The editors may ask for revisions of accepted manuscripts. Papers should have an abstract and key words. Please limit your paper submission to 8000 words, including footnotes and references, and format it for blind review (the text should be free of personal and institutional information). Along with the article, but not in the article, please send contact details: current affiliation, address, telephone number and email address. Authors are responsible for reinserting self identifying citations and references when manuscripts are prepared for final submission. Book reviews should have no more than 4000 words.

Formatting: All manuscripts should be formatted in Rich Text Format file (*rtf) or Microsoft Word document (*doc) with 12 pt. font size and double-spaced.

Documentation style: Please use the author-date system documentation style, as it is presented in *The Chicago Manual of Style*, 15th edition. For details please go to: http://www.publicreason.ro/submission-info

Notes: The contributions have to be original and not published before. *Public Reason* does not require exclusivity in submission of articles. However, if you have submitted your article to another journal, we kindly require that you let us know about it, and notify us without any delay about other journals' decisions about your article. The work has to be approved by all co-authors. Authors wishing to use text passages, figures, tables etc. that have already been published elsewhere are required to obtain permission from the copyright owner(s). Evidence that permissions have been granted are to be sent along with the manuscripts submitted for the blind review.The authors assume all responsibility for the content published in the journal.

Book reviews: Books for review should be sent to: CSRB, Department of Philosophy, University of Bucharest, 204 Splaiul Independentei, Sect. 6, 060024, Bucharest, Romania.

Advertising Information: If you want to advertise your books, department programs, conferences, events etc. please contact us at office@publicreason.ro in order to get general information or information on costs, specifications, and deadlines.

Subscription and delivery: *Public Reason* is an open access e-journal. For the print version of the journal please contact us at brancoveanu@publicreason.ro.

Online publication: *Public Reason* is available online at **http://publicreason.ro**

The journal *Public Reason* is published by *The Center for the Study of Rationality and Beliefs*. CSRB is a unit of scientific research within the Faculty of Philosophy of the University of Bucharest (http://www.ub-filosofie.ro). CSRB is located in the Faculty of Philosophy, Splaiul Independenței 204, Sector 6, postcode 060024, Bucharest. For more information see http://www.csrc.ro/EN/home

ISSN 2065-7285

EISSN 2065-8958

http://publicreason.ro.

ISSN 2065-7285
EISSN 2065-8958