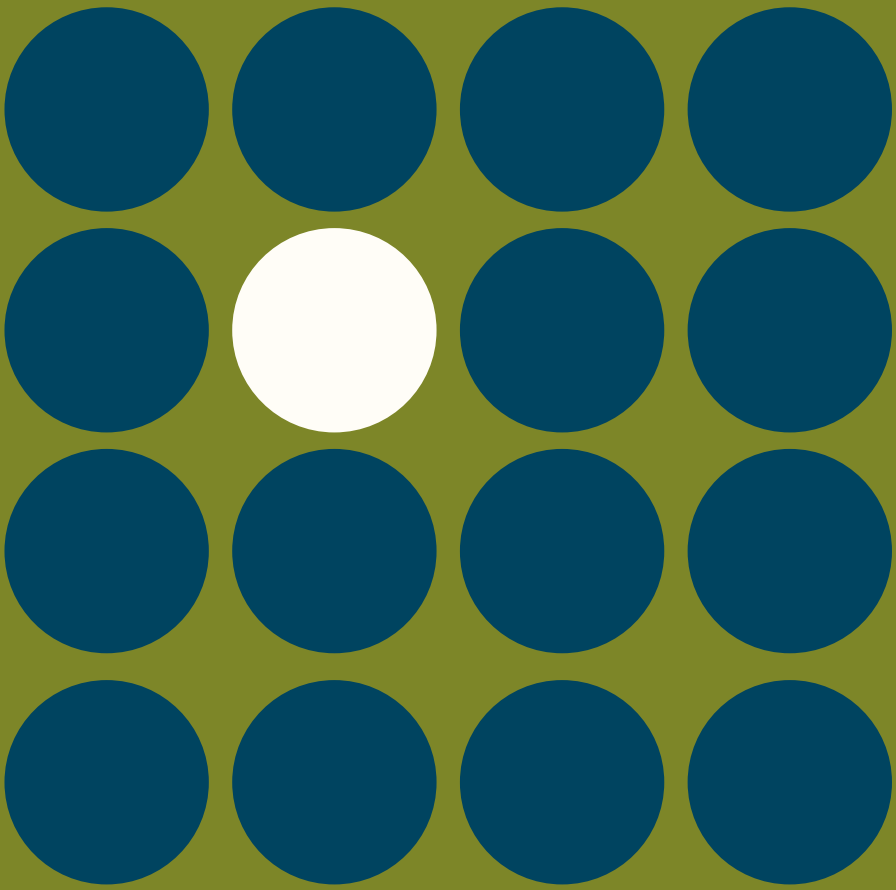
MORALITY, GOVERNANCE, AND SOCIAL INSTITUTIONS

REFLECTIONS ON RUSSELL HARDIN



Edited by
Thomas Christiano, Ingrid Creppell and Jack Knight

Morality, Governance, and Social Institutions

Thomas Christiano • Ingrid Creppell •
Jack Knight
Editors

Morality, Governance, and Social Institutions

Reflections on Russell Hardin

palgrave
macmillan

Editors
Thomas Christiano
Philosophy
University of Arizona
Tucson, Arizona
USA

Ingrid Creppell
Political Science
George Washington University
Washington, District of Columbia
USA

Jack Knight
Law
Duke University
Durham, North Carolina
USA

ISBN 978-3-319-61069-6 ISBN 978-3-319-61070-2 (eBook)
DOI 10.1007/978-3-319-61070-2

Library of Congress Control Number: 2017953938

Printed on acid-free paper

This Palgrave Macmillan imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgments

In November 2015, we hosted a festschrift for Russell Hardin at New York University. The conference brought together colleagues and former and current students from across the world to celebrate Russell—his path-breaking work, scholarly influence, and enormously generous contribution to the fields of political science, philosophy, and public policy. We wish to thank Diana Barnes, Director of Administration at the Wilf Family Department of Politics at NYU, who masterfully organized the complex logistics of the two-day proceedings. It would not have been possible to manage this without her help, conveyed with great efficiency and kindness. We also thank Melissa Schwartzberg, a member of the NYU faculty, who generously and graciously provided needed support on that occasion. Our greatest thanks go to Andrea Belag. Her loving support for Russell and for this celebration of his life and work were indispensable in the success of the conference. She welcomed dozens of visitors to their home during the event. We also want to thank the Wilf Family Department of Politics at NYU and in particular David Stasavage for their financial support.

In working on this edited volume, we received excellent editorial assistance from Chris Robinson and John Stegner of Palgrave Macmillan and Sam Bagg from Duke University. We are grateful to have been able to celebrate Russell and the work he has inspired in person and continue that recognition through the present volume.

Contents

Contributors

Michael Baurmann Heinrich-Heine-University Dusseldorf, Düsseldorf, Germany

Gregor Betz Karlsruhe Institute of Technology, Karlsruhe, Germany

Cristina Bicchieri Philosophy Department, University of Pennsylvania, Philadelphia, PA, USA

Geoffrey Brennan School of Philosophy, The Australian National University, Canberra, ACT, Australia

Thomas Christiano Philosophy Department, University of Arizona, Tucson, AZ, USA

Rainer Cramm Abt. Mikrobiologie, Institut für Biologie, Berlin, Germany

Ingrid Creppell Department of Political Science, George Washington University, Washington, DC, USA

Carolina Curvale Department of Political Science, FLACSO Ecuadors, Quito, Ecuador

Jon Elster Department of Political Science, Columbia University, New York, NY, USA

Gerald Gaus Philosophy Department, University of Arizona, Tucson, AZ, USA

Jack Knight School of Law, Duke University, Durham, NC, USA

Bernd Lahno Frankfurt School of Finance and Management, Frankfurt am Main, Germany

Paul-Aarons Ngomo Department of Politics, New York University, New York, NY, USA

Pasquale Pasquino Department of Politics, New York University, New York, NY, USA

Andrea Pozas-Loyo Instituto de Investigaciones Jurídicas, Universidad Nacional Autónoma de México, México City, Mexico

Andrew Sabl University of Toronto, Toronto, ON, Canada

Kimberly Stanton Tom Lantos Human Rights Commission, House Committee on Foreign Affairs, Washington, DC, USA

Erik Thulin University of Pennsylvania, Philadelphia, PA, USA

List of Figures

Chapter 8

Chapter 10

Chapter 13

List of Tables

Introduction

Thomas Christiano, Ingrid Creppell, and Jack Knight

Russell Hardin produced a body of work of great breadth and richness on essential subjects of the social sciences and political and moral philosophy: collective action, trust, utilitarian ethics, groups and conflict, institutions, and knowledge. The volume of output, the engagement with cross-cutting fields of scholarship and myriad subjects, and his at-times conversational mode of analysis make a succinct encapsulation difficult. We hope this volume will enhance appreciation for the power and analytical tools embedded in Hardin's work. His theoretical insights remain more applicable than ever. The most obvious thread running throughout his work is a rational choice approach to analyzing theory, policy, practices, beliefs, and events. Whereas some may consider this a procrustean framework, it served in fact to open inspection on the vast areas of ambiguity and indeterminacy in

T. Christiano (✉)
Philosophy Department, University of Arizona, Tucson, AZ, USA

I. Creppell
Department of Political Science, George Washington University, Washington, DC, USA

J. Knight
School of Law, Duke University, Durham, NC, USA

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_1

political and social life. Hardin brought a brilliant and incisive scalpel to investigating the extent and limits of self-interest as an explanatory variable and a human motive. He titled one of his books *Morality within the Limits of Reason* (1988), but his work has consistently proved rationality within the limits of the world—past, present, and future—a world individuals and groups continually make and remake within the constraints of resources, inheritances, and time.

In what follows we give a brief account of three main areas of Hardin's work: his distinctive take on the moral and political philosophy of utilitarianism, his accounts of collective action and the nature of social life, and his account of the fundamental idea of trust and social capital.

Moral and Political Philosophy

Russell Hardin's contributions to moral and political philosophy are animated by the desire to bring the social sciences and moral and political philosophy together. His most frequent criticism of much of contemporary moral philosophy is that it had become unmoored from social science during the twentieth century. Hume is his main inspiration here. The guiding moral theory for him is utilitarianism, which is elaborated in his *Morality within the Limits of Reason* (1988). The underlying idea is to start from a stark budget of moral ideas: the idea of the value of welfare, which is what people generally desire, and the idea that one ought to try to bring about welfare as much as possible. The rest of moral and political theory is generally concerned with means-ends reasoning concerning how to bring about welfare generally among persons using the results of social science. But even at the level of basic normative theory, Hardin thinks that economics shows us that a plausible conception of utilitarianism must be chastened by the idea that welfare itself is a poorly understood thing, that interpersonal comparisons of welfare are very difficult to discern in many important cases, and that aggregation of welfares across persons is poorly understood. As a result, what state of affairs realizes the greatest amount of welfare will in many cases be unknowable and consequently indeterminate. Hence what one must do in order to pursue the greatest amount of utility cannot be given a clear statement in many circumstances. The utilitarian principle's main implications are that one must help more people rather than less, satisfy the most basic needs, and pursue mutual advantage. Sometimes one may use material goods as proxies for welfare, but not always. Beyond these heuristics, it is hard to determine the extent to which an action or

institution constitutes an improvement in many persons' welfares over some other action or institution. A large area of indeterminacy remains.

Another aspect of his critique of contemporary moral philosophy is Hardin's lack of sympathy with the method of much of contemporary moral philosophy in that he, like Mill and Sidgwick before him, is very skeptical about the probative value of individual intuitions about particular moral examples. On his view, these intuitions are merely reflections of the ordinary norms and rules, which develop in societies in order to advance human interests and which are inculcated in us from early in our lives. They are thus very strongly tied to particular situations and social milieus. They have no further value. He reserves particular scorn for the efforts of some philosophers to attempt to test moral theories by means of intuitions about extraordinary and hypothetical circumstances. He thinks of these as reflecting merely the idiosyncratic sentiments of the particular philosophers. And these kinds of intuitive tests simply misuse the sentiments by employing them outside of their appropriate sphere of application.

Hardin does not think that philosophy can get by without any intuitions at all. He expresses confidence in universal intuitions such as the value of welfare and the importance of bringing about as much welfare as possible. He admires Kant's theoretical effort to derive morality from a small set of fundamental intuitions, though he rejects the categorical imperative as unsuited to the evaluation of action given the strategic interaction of agents with each other. The consequence of this is that he regards particular moral rules and institutional structures as justified to the extent that they bring about human welfare in the circumstances.

Hardin's moral and political theorizing is guided by the idea that human beings are primarily motivated by self-interest and only occasionally altruism, like Hume and Bentham. He takes this as fundamental because he thinks that there is good evidence in favor of the thesis that one can explain how human beings act and how institutional structures function by invoking self-interest alone, evidence which we discuss in more detail below. He employs and develops a highly sophisticated rational choice approach to the explanation of human action and the development and operation of institutions. The main tools he uses are game theory for small and large numbers and an economic theory of information.

For all that, the basic account of utilitarianism is act utilitarian. Our actions are hemmed in by low levels of information and by self-interest. They take place in circumstances in which we strategically interact with others and rely on information that other people possess. So Hardin argues

that we need to construct and mainly to maintain institutions to provide the right strategic background against which to act. Mostly the work we do on institutions is to preserve them but we occasionally attempt to change them. The limits to information are such that our abilities to change institutions in the ways we want are themselves quite limited. Not only are we uncertain about the effects of bringing about institutions, there are costs to bringing them about.

The central game theoretic ideas for Hardin in the explanation of institutions are the ideas of coordination and the prisoner's dilemma. Coordination plays this central role because institutions and norms that are based on coordination are self-enforcing in a way that other institutions are not. That is, if we suppose that human beings are primarily self-interested, the only institutions that can guarantee that we act together in mutually fruitful ways are ones that are based in coordination.

The prisoner's dilemma plays a large role in explaining the failure of contractarianism as well as the justification of the institutions of property and contract. Contractarianism fails because it relies on the idea that persons will act to do their fair share even when it is not in their individual interests. Self-interest will not sustain a large-scale social contract because there is a large number prisoner's dilemma for each self-interested individual. The idea that people will uphold institutions because they act from a sense of justice, as Rawls maintains, is entirely foreign to Hardin's conception of institutions.

Hardin is skeptical of much of what we describe as ideal theory in moral and political philosophy. He maintains throughout his work a strong skepticism about such contemporary political ideals such as the equal distribution of goods or opportunities, democracy, or distribution in accord with desert. He is very skeptical about the nature of the achievement of John Rawls' theory of justice and contractarian thinking in general. Some of these ideals he considers counterproductive under current circumstances while others are simply impossible to achieve. And these theories are generally marred by the fact that they fail to take into account how we are to get from where we are to their supposedly ideal state. This gives his political theorizing a distinctly incrementalist and skeptical character. Also, like Hume, he tends to theorize politically more by determining which political institutions work to promote welfare and how they work.

The main subject of normative thinking for Hardin is the justification of institutions. The foundation of Hardin's moral and political thought is his concern for constitutions. The state, on his account, is necessary to provide

the background for beneficial strategic interaction especially among strangers. Constitutions are potentially successful political institutions to the extent that they serve as coordination points for primarily self-interested individuals. They can achieve a certain degree of stability for the society when they coordinate. Thus, they serve a highly useful role in organizing societies in particular in averting violent conflict and setting the framework for mutually beneficial social interaction. Hardin's discussion here builds on an historical and social scientific analysis of how constitutions work, when they do work. Hardin is an ardent defender of liberal rights of privacy and of property and contract. He thinks they are institutions that enable people to pursue mutual advantage without any commitment to altruism and despite the severe limitations on information that people have in pursuing their interests. Hardin thinks that the basis of rights consists in their being institutional devices that enable us to advance our welfares on our own and enable us to engage in mutually beneficial arrangements with small numbers of people when each has the best appreciation of the effects of actions on her own welfare. The rights of privacy, property, and contract are all derivative from the utilitarian principle and the facts of limited information and self-interest. His defenses of property and contract are completely instrumentalist. Though he is an admirer of F. A. Hayek, he rejects what Hayek called the classical liberal approach to property and contract. There are many circumstances of strategic interaction in which these rights ought to be modified. For example, Hardin argues that limitations of individual rights of contract are defensible when there is a need for collective protection of groups of persons against the possibility of free riders who contract individually. The cases of the protection of unions or the forbidding of vote selling are good illustrations of this kind of solution to a strategic problem for Hardin.

There is some room for considerations of distributive justice on Hardin's account as in classical utilitarian accounts. This is largely because he views the economic product of society as mainly a function of how the society is organized and does not tie the wealth of persons to their own efforts. He agrees with Arrow and Rawls that the productivity of each person is mainly tied to the way that the surrounding society complements that person's efforts and capacities. There is little room for desert in his view or any view that ties a person's product directly to that person. Furthermore, he thinks that the modern state, unlike its eighteenth century predecessor, is capable of significant redistribution. Yet these considerations are hedged in by the fact that only crude interpersonal comparisons of welfare are typically

justified and that inequality may be necessary to supply incentives to action for the more talented. Furthermore, modern societies have not shown much will to redistribution.

This latter observation is connected to Hardin's generally very critical approach to democracy, fully developed in his *Liberalism, Constitutionalism, and Democracy* (1999). Here Hardin generally follows the thought that citizens in democracy generally have little or no incentive to become informed about politics because they are primarily self-interested. When one multiplies the value of the effect of the outcome of voting on one's interests by the extremely small probability that it will make a difference, and one considers the real costs of becoming informed, we can see that the marginal cost of gaining new information will be greater than the marginal benefit. The rational citizens will likely not be even moderately well informed. But a society run by people who do not have any knowledge about society is not really run by citizens at all. Furthermore, Hardin thinks that the results of social choice theory imply that collective decision rules are likely to issue in decisions that fail to satisfy certain elementary requirements of consistency. The ideals of popular rule or political equality consequently hold little sway in his work.

Yet Hardin is by no means an opponent of democracy. Given his adherence to the self-interest analysis of human action, the dangers of oligarchy are very clear. He argues that democracy is a bulwark against oligarchic rule. Hardin's attitude to democracy is ambivalent here. He thinks that it has failed to achieve significant redistribution, which would be desirable from a utilitarian standpoint, because of the large-scale prisoner's dilemma democracy poses. But he also argues that the lack of serious participation in politics of most people is often beneficial since it tends to tamp down the level of conflict in democracies.

Overall, Hardin makes a powerful contribution to moral and political thought by continuing the project of a utilitarian conception of morality with the help of the new tools of contemporary social and economic thought. As one would expect, this is a work in progress. But Hardin's work has transformed the project in a deeply illuminating and far reaching way, as demonstrated by the contributors in this volume.

On morality, Gerald Gaus takes up the challenge of Hardin's analysis of cooperation as premised on self-interest-based conventions and argues for a conception of social morality guided by internalized "social-moral rules." Bernd Lahno, considering Hardin's extension of Hume's moral and political theory, argues that some version of freedom (Constant's "liberty of the

ancients") may not be totally inaccessible in the face of Hardinian pessimism. Ngomo discusses Hardin's account of Hobbes's political thought. Others such as Geoff Brennan and Cristina Bicchieri flesh out the promise of Hardin's conception of norms as effective because of their dual power to locate self-interest and tap into moral psychology.

On constitutions and constitutionalism, Hardin's integration of social science and constitutions as conventions (as opposed to contracts), influenced many scholars, and should provoke many more, as Andrew Sabl persuasively argues in his essay, about the paradigm-challenging nature of Hardin's work, which also tended to make it more difficult to fit it into the conventional norms of scholarship. Pasquale Pasquino sets forth a theory of constitutional democracy underscoring the preeminence of courts over legislatures, a position Hardin *might* support based on utilitarian reasoning about which political institutions would tend to protect mutual advantage better over time. Advancing Hardin's persistent aim to think through the consequences of various types of constraints, Andrea Pozas-Loyo investigates the relative effects of constitutional law versus social norms in her study of a Mexican president's decision to seek re-election.

Rationality, Collective Action and Community

Hardin's contributions to social science complement his moral and political philosophy. Hardin's first book *Collective Action* (1982) addresses the relationship between individual rationality and group action. From the standpoint of narrow rationality, participating in collective action will often appear not to be in an individual's self-interest. When and why does participation take place then? Examples of successful group action abound but failures to coalesce and act for group goods may be even more pervasive. Hardin seeks explanation for both provision and failure by using a rational choice framework. By "rational" he means "efficient in securing one's self-interest." The virtues of this approach are certainly methodological: one can take apart opaque phenomena and show comprehensible and plausible explanations for how things turn out in the way they do. His work in this early text indicates an abiding interest in the forces that keep good public policy from advancing. He takes the domain of collective action to encompass actions that we would consider beneficial: the elimination of "bads" such as polluted air and water, and the provision of goods, like participation in civil and women's rights movements, voting, support for water irrigation systems, environmental conservation, and so forth. He keenly notes "how little Americans have spent on such honored causes" as environmentalism,

gun control, and so forth. When they have done so, it is due to structural features of a situation (e.g., conditions allowing leaders to emerge), to organizational factors, and to the activation of extra-rational motivations (morality, desire to participate in history, misunderstanding/ignorance). Achieving rational collective ends requires indirect and structural features of social and political life to make these ends dovetail with the self-interest of individuals. Beginning from the point of view of the individual decider helps analysts to clarify "the impact of other motivations" (he insists: "many social and individual phenomena cannot readily be explained as the product of interest-seeking by individuals" 2015, 898). Presuming the collective goods provided, we might hope to re-construct situations of choice to maximize these goods.

More than a decade later, in *One for All, The Logic of Group Conflict* (1995), Hardin takes up the question of group identity and collective "bads" of a different sort—entrenched animosities, genocide, dueling, vendettas, and so forth. Inspired to understand "the sway of groups in our time"—a period after the collapse of the Soviet Union with the explosion of ethnic war in Eastern Europe, genocide in Rwanda, Québécois nationalism, among others—he again aims to show the large part played by self-interest, but now with regard to apparently self-defeating and irrational behaviors. In the book, he confronts primordialists, theorists who ascribe conflict to the resurgence of deep-rooted permanent antipathies, and moralists who point to the grip of moral imperatives. Hardin argues for the rationality of individual behavior operating in these perplexing cases. This work, and subsequent analysis in *Indeterminacy and Society* (2003) and *How Do You Know?* (2009), toggles among the individual level of decision-making, the context of choice, and the collective level of large-scale phenomena to trace out connections leading to "grossly harmful effects" and disastrous consequences in particular situations. *One for All* delves into rich details of history and politics to explain the relative balance of factors—individual self-interest and situational—affecting the sway of groups. We briefly consider three insights from Hardin's analysis in this work: (1) identity as identification, (2) the grip of norms, and (3) descent into violence.

Identification

The question of identity is a central organizing subject in *One for All*. The subject of "identity" pervades the social sciences yet remains without a single canonical literature to organize arguments. Erik Erikson's work

may be closest to a core reference. Hardin agrees with Erikson's approach: "the central problem of identity is identification, what motivates you, not what characteristics you have" (xi). Off the bat, Hardin rejects the approach to identity as a static set of features, which objectively define a set of persons or which people carry around in themselves as a picture of who they are (with corresponding motivations to be read off the ascribed identity). His insistence on identification presents the problem of identity as one about an active, ongoing process of choosing what an individual will do and be in various situations. The question in *One for All* focuses specifically on *group* identification. Hardin sees identification as the active engagement with "quasi-objective identities" pervading options in the social world, and most importantly the commitments a person adopts to acting as a member of a particular group.

The connection between self-interest and group identification is at the core of the account. If membership cannot be objectively determined but depends on people's acceptance of an alignment, then we must ask why people accept and decide to abide by certain defining characteristics. Rationality enters into identification not in the simplistic sense that it is "rational to adopt a particular identification with its associated beliefs." Hardin insists this is "patently false and beside the point." Rather, he explains, "it may be rational to do what produces a particular identification and, once one has that identification, it is commonly rational to further the interests determined by that identification" (OA, 60).

The necessity of group-based behavior is rooted in the evolutionary advantages of coordinating human order through group differentiation. Hardin argues that once coordination around a particular defined collective happens, individuals find themselves associated and the ability to be part of a coordinated unit brings advantages of two main sorts: (1) concrete benefits the group provides to security, access to resources, jobs, and so forth and (2) pleasure derived from membership in a group, as intrinsic well-being is enhanced through building and sharing cultural values, what Hardin calls the "comforts of home." In his terminology, the linkages might be simplified as follows: coordination (among group members) → advantages: access to resources, common expectations, and epistemological benefits (comforts of home) → moral valuations.

The most significant contribution of Hardin's theory of identification, we would argue, is that it operates as both an input and an outcome of action. Whereas many theories treat identity as a causal input in an explanation of violence or some event, for instance, Irish Protestants and

Catholics participating in the "troubles" in Northern Ireland, Hardin insists on a more active understanding of identification as enacting one's categorization (for purposes of protection and associated benefits) rather than just protecting one's "being." Thus, the motivations of such groups cannot be reduced to an objective identity—Catholic versus Protestant—but must be analyzed as a dynamic that involves self-interest of persons given particular options and aims, and the larger-scale exogenous forces that constrain choice. This he calls a method of deconstructionism and contextualism, which demonstrates the pressures of the fulcrum of deciding to act, arising from the individual's position in a context of factors, internal and external, short-term and long-term. Rationality operates insofar as within the eddy of forces, the individual must strategically decide what will most conduce to her benefit, as this person presently situated. Thus, identifications are pre-given options but also essentially rooted in the self-interest of persons to survive and thrive, vis-à-vis other groups; the identifications must therefore continually be enacted and reinforced.

Exclusionary Norms

Essential to Hardin's analysis of the work of groups is his description of the purpose and role of norms of difference and exclusion. In order to keep persons in line with the coordinated power of the group—to ensure self-interest remains clearly linked to the group's welfare and comparative position—conventions and more quasi-formal norms of exclusion enforce commitment.

Hardin explains the force of rationality not only at the individual level of commitment but also at the macro-level in his explanation of the functional benefits of norms, by which identifications are enacted, maintained, and enforced on members. Here he takes the task not to explain the initial appearance of a potential convention of behavior marking off a group. To do that one would have to "investigate millions of actions by vast numbers of people over several years." Rather, once a convention emerges and takes hold (and he uses the adverb once quite frequently in his analysis)—for example, dueling as a mark of aristocracy—individual members become compelled to submit to this exclusionary, gate-keeping norm because the failure to do so will lead to ejection from the group. In acting according to the norm, the power and success of the group are proven by the people who must themselves be subject to it. Through obeying they infuse the rules with power and thereby the ballast of their group. Thus, the maintenance and

propagation (though not emergence) of general norms of difference and exclusion are demonstrated to conduce to the self-interest of group members because those norms function to sustain the group, which in turn brings benefits to individuals within it.

Hardin compares the naturally divisive tendencies of norms of exclusion to universalistic norms like promise-keeping, truth-telling, and loyalty among associates. The former contribute to identifications and serve as a guide in political–social mapping. Their hold seems more obviously connected to a self-interest dynamic through the reasoning of in-group/out-group. Universalistic norms appear to have a weaker grip, but again self-interest comes into play in dyadic, ongoing relationships, when one expects to interact over time. To the question of how nondyadic universalistic norms can prevail, Hardin sees them operating again at the level of groups. Here, for instance, the practices of vendetta are interpreted as universal norms of loyalty deployed for one's particular group against another.

Hardin helps the perplexed understand why persons who may not always agree with the content of norms will go along with them. For instance, on racism, he observed: "Even those who were not racially prejudiced therefore may have participated in racial discrimination—because it seemed costly not to do so" (OA, 90). His studies reveal the purchase of norms—individually comprehensible and at times apparently compulsory. Many will bring the comforts of home and benefits of belonging, language rules being a prominent example. Yet, they also carry socially harmful, abhorrent, and at times disastrous consequences.

Violence: Inputs and Outcomes

Perhaps the most innovative aspect of Hardin's theory of groups comes in at this point. Were norms of difference and exclusion to function merely to maintain distinctions and therefore in-group/out-group ordering, we would accept the basic rationality of the system. Yet, group ordering produces conflict, competition and can escalate into violence. When social order becomes destructive, we ask how it will be in someone's rational interest to support violent actions and outcomes. The question is not whether it will be in someone's interest to support violent action once a dynamic of tit-for-tat takes off; self-preservation would demand one do so. Rather, what would make an extreme practice like genocide, dueling or destruction of the Yugoslavian state a more rational path to take than a less extreme form of seeking superiority or even domination? Hardin insists

that people's aims do not transform into monstrous desires to enact collective horror; rather normal self-interest-based decisions may end up here.

Dueling typifies the logic of a norm of exclusion leading toward violence. As a practice, dueling channels the public nature of aristocratic status: a gentleman's position required public displays of nobility. Were one's personal honor to be insulted, one must demonstrate bravery in a visible test of standing through marksmanship. Publicity, manly honor, stylized deployment of a deadly weapon, and courageous risk of death—one can see how the elements of the duel might have come together. Once this coalesced into a group norm, members of the class would have been compelled to abide by it if they sought to continue enjoying advantages of the rank. Hardin's investigation into this seemingly perverse norm makes comprehensible how aristocratic persons born into such a rank, or persons seeking to prove belonging, would have enacted such a strange honor ritual.

Hardin also returns repeatedly to the Yugoslavian descent into ethnic war in the early 1990s. A self-interest-based explanation must do more than re-describe the battles as driven by groups fighting for their own interest, which amounts to a mere repackaging of the event into rational choice terminology. Hardin zeros in on the indirect and cumulative forces which would have led to a situation in which individual motivations end up leading from ethnic conflict to ethnic violence, the latter a qualitatively different phenomenon. Hardin pointedly denies attributing bloodthirsty desires to the protagonists. Normal perception, fear, and incentive explain the escalation through the following logic: (1) a period of political transition in which the "tottering weakness of the central regime" is combined with (2) vociferous Serbian nationalism provokes a mirroring response by Croatian nationalists and accompanying nationalist movement. These lead to alienated segments of Serb minorities in Croatia taking up paramilitary actions and igniting a descent into violence. Hardin describes the effect: "Violence is a tipping phenomenon because, once it begins or reaches a high enough level, it is often self-reinforcing. Violence can provoke reprisals and preemptive attacks" (OA, 155). The important point for grasping the tragic unfolding is that people of goodwill are "forced to choose in some of these moments" and that they may panic in responding because of a perception of the other's aggression and the options available to protect themselves and, for the leaders, their people. Leaders often play an outsized role in determining the unfortunate fates of their followers.

Hardin insists in his analysis of genocide, dueling, and other assorted lethal collective action that we begin from the point of view of persons

seeking normal self-preserving interests. Hardin demonstrates how through indirect accretion of conventions, and group consolidation, points of action and decision can be reached at which normal people will tip over into genocidal monsters. One does not need to be a primordialist or a communitarian to understand the emotions, ideas, beliefs, and incentives that can prompt these behaviors. Indeed primordialists and communitarians get it wrong when they see the source of lethal action as the inclination of people to harbor deep-rooted bonds of unity and antipathy. In contrast, Hardin argues for the constant theoretical application of the assumptions that people act to preserve themselves, and decide within the limits of their perceptual lenses and the limits of action on the available "stages" or spaces of enactment.

Hardin's work in *One for All* sets the stage for his subsequent exploration of indeterminacy in *Indeterminacy and Society* (2003) and of what he called commonsense epistemology in *How Do You Know? The Economics of Ordinary Knowledge* (2009). This pair of books was written to address two angles on the problem of the limits of rational action: "the individual's capacity for achieving objectively good outcomes is often impaired or even stymied" (2009, xii) due to problems of knowledge and of strategic interaction. Both of these theoretical trajectories can be found in *One for All*, where group identification was examined as an additional complex reality for individual motivation and collective outcomes.

Out of many incisive discussions, one in particular seems to capture the essential dimension of his conception of the human situation: the connection between knowledge, identity, and action. He writes: "What it is rational to do depends on who one is, that is what knowledge one has" (OA, 17). Identification depends on knowledge, and knowledge is always truncated and limited by past and present blinders, narratives and interests. As he memorably observes: "Our sunk costs are us. Our cultural sunk costs have been transmuted into information and putative knowledge that is not merely gone. Much of it is a resource to us in our further actions—although much of it is perhaps an unfortunate resource, more nearly an obstacle, and we might wish it were gone" (OA, 69). Is there a way out of this dead-end? The point of Hardin's work is to seek remedies for the limitations—epistemological and strategic—which misdirect human thinking, feeling and acting on a collective scale. We might invoke one of his heroes, Thomas Hobbes, for a reminder of the point of collective knowledge and the work of politics. Understanding our own conditions and myopia should provide grounds for constructing political institutions to solve some of the dismal

results of uncoordinated, suboptimal and at times lethal interactions. Political rules then can productively enhance the fortunes of most people, who for Hardin, are not misanthropes and do not desire to dominate and plunder their fellow citizens. "For Hobbes, coercion is necessary to prevent the few who may be ill-intentioned from harming the many who are well-intentioned. Even more important . . . the possibility of sanction is valuable for letting the well-intentioned, who do not require sanctions, risk being cooperative on the secure knowledge that those with whom they come to interact are similarly well-intentioned" (1981, 185–186). Thus, we see how Hardin's own life-work, the expansion of the storehouse of collective knowledge, is carried out in the hope of advancing human welfare, as well as to increase self-understanding.

Three chapters in this volume directly spring from an application of Hardin's theory of group conflict. Baurmann et al. investigate the origins of extremist ideologies through a model of dynamic belief formation that unfolds through mutual adaptation, supporting Hardin's analysis of the indirect and yet seemingly inexorable outcomes of collective radicalization. Carolina Curvale offers a rational choice account of ethnic conflict in Northern Ireland and the strategies of deployment and disarmament in the cases of the Irish Republican Army (IRA) and Sinn Fein. Kimberly Stanton considers Hardin's theory of human rights as a case of universalistic norms, which can only compel weakly, but may carry additional modes of institutional incentives.

Trust as an Explanation of Social Cooperation

Hardin's long-standing interest in explanations of successful collective action leads him to undertake an extensive analysis of the concept of trust. Trust and social capital are related concepts that have emerged as rich sources of social scientific analysis of how social cooperation occurs in a world characterized by the logic of collective action. In this section, we highlight two major features of Hardin's analysis of trust. First, we consider his substantive claims about trust and trustworthiness, emphasizing the distinctiveness of his approach relative to the dominant views in the literature. Second, we reflect on what his analysis of trust tells us more generally about his approach to social explanation.

To do so, we will begin with a brief sketch of the state of the research on trust in order properly to demonstrate the distinctiveness of Hardin's approach. Trust and its related concept, social capital, have been invoked to explain various forms of social cooperation in political and economic life.

Although, as with many such ideas, those who invoke it do not all share the same meaning of the concept, they do, with few exceptions, posit trust as an unquestioned good, a necessary condition for a healthy and productive society.

When Hardin began his analysis of trust, the concept was being used to explain cooperation in a wide variety of social situations. From the most common, the two-person interaction to achieve a common goal or purpose, to the most general, the willingness of an individual to contribute to the production of a public good, trust was offered as an explanation of successful cooperation. In the social sciences, there was an understandable focus on the societal level, on a general sense of trust in the other members of one's society. This is the form of trust most closely related to other concepts like social capital. Its most important features are that it is generalized across the members of a group (in the sense that it is not tied to the reputation of particular individuals) and that it is informal in nature (in the sense that its causal effect is not related to any formal sanctioning mechanisms like the state).

A fundamental problem for this research was that it is difficult to identify the specific mechanisms by which trust might facilitate social cooperation. This follows from the fact that the phenomenon of trust, as conceptualized in the research, is so intertwined with the cooperative behavior that researchers are trying to explain that it is difficult to sort out what is actually doing the work in the explanation. Thus, the question, to what extent is trust an independent factor in fostering cooperation?

The intuition behind this research is that trust might be an explanation for cooperation in those circumstances in which we are uncertain about the likelihood that others will cooperate with us. Several accounts have been offered as an answer to this question. In his analysis of the existing literature, Hardin categorizes the alternatives in three groups, in terms of the basic causal mechanism on which the account is based: cognitive expectations and beliefs, personal dispositions and moral commitments (Hardin 2006). The first is the account most consistent with the logic of collective action, seeking to identify mechanisms that would enhance expectations about cooperation to such an extent that rational, self-interested decision-making would recommend cooperative behavior. Gambetta's characterization captures the dominant conception: trust as "a particular level of subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action (or independently of his capacity ever to be able to monitor it) and in

a context in which it affects his own action" (1988, 217). With this characterization, the key to the explanatory question became, what factors establish the particular level of subjective probability necessary to achieve trust?

The other two accounts are grounded in noncognitive mechanisms, seeking explanations in terms of psychological motivations or moral commitments. On these accounts the identification of trust involves a search for the factors that explain why some people are more willing than others to believe that their acts of cooperation will be met in kind.

Each of these accounts of trust is used to explain cooperation both in the two-person and the general case. For the noncognitive accounts, the transition of the explanation from micro to macro interactions is fairly straightforward. The psychological motivations and/or the moral commitments that explain small-*n* cooperation are basically the same as those that explain generalized trust. For the cognitive accounts, this transition is more complicated. Put in terms of social expectations, theories of generalized trust must provide an account of how these expectations of the positive likelihood of cooperation extend beyond the instances of interactions with those whom we actually know to a more general expectation of cooperation about the members of a society. This requires an explanation of how the individual members of a society acquire these shared beliefs. The research focused on two primary sources of expectations and beliefs: (1) individual past experience and (2) cultural factors. While past experience is a major source of belief formation for each individual, the diversity of experiences makes it unlikely that it will be the primary source of generalized trust.

Hardin's analysis of trust builds on his earlier work on collective action. The logic of collective action suggests that if trust is an independent factor in explaining cooperation, it would somehow involve the cognitive relationship between expectations and the choice to cooperate. Hardin adopts several initial premises that follow from his previous work. First, he develops his own account at the micro level, analyzing the two-person interactions that he feels characterize most of our day-to-day experiences. As in much of his work in this area, he argues that if trust matters for cooperation, we should be able to identify the relevant mechanism at the most basic level of social interaction. Second, he insists that the basic concept of trust should be the same regardless of the level of analysis. It should be applicable to small-*n* relationships with people whom we know well and also to the kinds of large-*n* interactions that characterize the generalized trust research. For Hardin it follows that if the conception that offers the best explanation in small-*n* cases cannot explain the causal effects of trust in the generalized case, then it

raises serious questions about the applicability of trust as an explanatory factor in larger social interactions. Third, he argues that we should keep an open mind about the value of any such causal mechanism, questioning the dominant perspective that trust is an unquestioned good.

From this perspective Hardin develops his encapsulated-interest conception of trust (Hardin 2001). This account challenged all of the existing conceptions of trust. He conceptualized trust as a three-part relation: "A trusts B to do X." For Hardin, trust is a dyadic conception that always refers to particular actions. He rejects the open-ended conception that "A trusts B." According to the encapsulated-interest account I trust someone when I believe that he or she has some reason to act in my best interests or at least to take my best interests fully into account. Hardin envisions this conception as a significant improvement on other cognitive accounts grounded in the logic of collective action: "my trust in you is encapsulated in your interest in fulfilling the trust. It is this fact that makes my trust more than merely expectations about your behavior. My expectations are grounded in an understanding (perhaps mistaken) of your interests specifically with respect to me" (2001, 3).

The thrust of this proposal is to identify the mechanism by which trust might influence social cooperation in the interests of the trusted party. Hardin insists that trust is more than mere expectations about future cooperative behavior, as the dominant trust accounts within the collective action literature would have it.

The interesting question then becomes, what is the basis for my beliefs about other people's interests? Hardin looks to the kinds of relationships that are characterized by reciprocal interests. The most obvious examples involve such relationships as those in which other people value my own welfare (e.g., family relationships) or situations in which the other person values the relationship with me (e.g., friendships or romantic relationships). On the encapsulated-interest account, the more valuable the relationship is to the parties involved, the more trusting we are likely to be.

Note that this conception rules out many situations that are treated as trust relationships in the literature. First, Hardin rejects the account of trust that was grounded in moral commitments. While not ruling out the possibility that some trust relationships might be based on the moral commitments that some individuals have to being trustworthy, he argues it is quite unlikely that such strong moral commitments form the basis for most ongoing trust relationships. He thinks such commitments would only serve to facilitate trust and thus cooperation in the narrow range of

interactions in which the parties know each other well. Second, Hardin rejects any account of trust that is grounded in noncognitive, dispositional motivations. In regard to these accounts, Hardin suggests that if some people trust others purely because of the disposition to do so, they have probably made a previous cognitive decision about whom they would trust in that way. He doubts the plausibility of any dispositional account that could not adequately distinguish between the contexts in which our levels of trust might vary.

With this conception Hardin is better able to identify the circumstances in which trust might be an independent factor in facilitating social cooperation. And in doing so he hopes to minimize some of the primary conceptual confusion in the literature. He is always concerned to distinguish trusting someone from acting on that trust. He notes it is not trusting someone that is risky but rather acting on that trust. On his account trust is about knowledge while acting on trust is about behavior. Thus, to say that one chooses to trust me mistakenly implies that trusting is a matter of action. Hardin argues that by keeping this distinction clear we can avoid the problem that bedeviled much of the research on these issues, of conflating trust as a causal belief with the act of cooperating itself.

With the encapsulated-interest conception, Hardin makes several important substantive contributions to the analysis of trust. One of the most important of these is his emphasis on the distinction between trust and trustworthiness (Hardin 2004). Referring to the basic trust relationship "A trust B to do X," trust involves A's belief about B's interest in the implications of X for A, while trustworthiness is a characteristic of B that relates to B's interests in regard to A. Hardin argues that much of the debate about trust is better understood as debate about trustworthiness. Much of the discussion about how to foster trust in society is on his account actually a discussion about how to foster trustworthiness.

Hardin thinks that one of the strongest recommendations for the encapsulated-interest account is that it helps to explain degrees of trust and distrust at every level of analysis (Cook et al. 2007). And it does so, on his account, without any conceptual changes. For the research on generalized social trust, this has important implications. Hardin questions the value of the concept of generalized or social trust. In regard to the research on generalized trust, he argues that it is not capturing evidence of trust but rather evidence of positive expectations of other's trustworthiness or cooperativeness. With his characteristic directness and clarity, Hardin rejects the concept of generalized trust as an explanation of social

cooperation: "In any real-world context, I trust some more than others, and I trust any given person more in some contexts than others. I may be more optimistic in my expectations of others' trustworthiness then you are, but apart from such a general fact, I do not have generalized trust. I might also typecast many people and suppose some of the types are very likely to be trustworthy and therefore worth the risk of cooperating with them, other types less so, and so others not at all. But such typecasting falls far short of generalized trust. It is merely optimism about certain others" (Hardin 2001, 14).

Hardin offers an important insight about the relationship between trust and trustworthiness, and in doing so helps explain why there has been so much conceptual confusion in the literature. The insight is simple and yet persuasive: trustworthiness commonly begets trust. If I act on your trustworthiness and am subsequently rewarded by your cooperation, it will tend to enhance my trust in you. He thinks that the close connection between the two concepts is one explanation for why researchers tend to conflate beliefs and actions in their analyses of trust.

Throughout his work on trust Hardin seeks to clarify types of circumstances and conditions under which beliefs about trust and trustworthiness can be justified. Invariably this involves an identification of the proximity of the interests of the parties involved as well as a clarification of the differences between trust and trustworthiness. For example, Hardin reassesses the large literature on the decline of trust in modern society and reinterprets it as a concern about the decline of perceived trustworthiness in those societies. In addition to the advantages of providing a more precise and persuasive explanation of the sources of social cooperation, Hardin believes that the clarification of the trust-trustworthiness conceptual relationship would serve to enhance practical policy efforts to create the conditions for greater social cooperation.

Hardin's justification for his encapsulated-interest conception provides a broader insight into how he envisions the intellectual task involved in social theory. He justifies the encapsulated-interest conception in terms of its ability to help us both explain and evaluate behavior. Hardin wants to bridge the gap between philosophical (which was primarily definitional and conceptual) and social scientific (which was primarily explanatory) work on trust. He envisions the encapsulated-interest conception as a concept that can satisfy both the definitional and the explanatory tasks. And yet he is not committed to any form of essentialist conception of trust, being clear to explain that he is not attempting to present "the" true meaning of trust.

As an example of the heuristic power of Hardin's expansion of the trust concept, Jon Elster's chapter in this volume examines collective action in America before 1787 and demonstrates the centrality of trust and distrust for the Federalists and for ordinary people attempting to create a coherent political group.

Hardin always adopts a very practical approach to social explanation. One interesting example is his approach to functional explanations. In assessing the merits of functional explanations of trust, he acknowledges that such explanations fail to satisfy the rigorous criteria that are established in the literature for a successful functional explanation (e.g., Elster 1979). And yet he finds that thinking about the ways in which trustworthiness is functionally beneficial to social cooperation illuminates a wide range of social situations in the world today. His focus is not on the formal criteria but rather on how the informal logic of functional benefits enhances our understanding of everyday life. For Hardin, the success of a social explanation rests primarily on the common sense ways in which it advances our understanding of the world.

By extending his analysis of the logic of collective action to incorporate how trust and trustworthiness might help to better explain social cooperation, Hardin asks to be judged by a very pragmatic criterion: "I put forward a workable notion that can be used to cover much of our experience of relying on others in that it can be used to help explain variations in our behavior and beliefs about the reliability of others, including collective others. My central concern is such explanation" (Hardin 2001, 9). His work on trust and trustworthiness well satisfies that goal.

Chapters in This Volume

We have grouped the chapters in this volume under three headings. The first set of chapters by Gerald Gaus, Geoffrey Brennan, Bernd Lahno, and Paul-Aarons Ngomo includes theoretical discussions of basic concepts in social science and political theory such as social morality, self-esteem, or historical figures such as Thomas Hobbes and Ancient philosophers. The second set of chapters by Andrew Sabl, Jon Elster, Pasquale Pasquino, and Andrea Pozas-Loyo is focused on basic normative and theoretical questions about constitutions as social institutions. The third set of chapters by Michael Baurmann/Gregor Betz/Rainer Cramm, Kimberly Stanton, Carolina Curvale and Cristina Bicchieri, and Erik Thulin is on issues of group formation and violence.

References

Cook, Karen, Russell Hardin, and Margaret Levi. 2007. *Cooperation Without Trust*. New York: Russell Sage Foundation.

Elster, Jon. 1979. *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

Gambetta, Diego. 1988. *Trust: Making and Breaking Cooperative Relations*. Oxford: Blackwell.

Hardin, Russell. 1982. *Collective Action*. Baltimore: Johns Hopkins University Press.

———. 1988. *Morality Within the Limits of Reason*. Chicago: University of Chicago Press.

———. 1995. *One for All: The Logic of Group Conflict*. Princeton: Princeton University Press.

———. 1999. *Liberalism, Constitutionalism and Democracy*. Oxford: Oxford University Press.

———. 2001. Conceptions and Explanations of Trust. In *Trust in Society*, ed. Karen Cook. New York: Russell Sage Foundation.

———. 2004. *Trust and Trustworthiness*. New York: Russell Sage Foundation.

———. 2006. *Trust*. New York: Polity Press.

———. 2007. *David Hume: Moral and Political Theorist*. Oxford: Oxford University Press.

———. 2009. *How Do You Know? The Economics of Ordinary Knowledge*. Princeton: Princeton University Press.

———. 2015. Rational Choice Explanation: Philosophical Aspects. In *International Encyclopedia of the Social & Behavioral Sciences*, vol. 19, 2nd ed. New York: Elsevier.

Thomas Hobbes, L., ed. 1981. *Introduction by C. B. MacPherson*. Harmondsworth: Penguin.

The Priority of Social Morality

Gerald Gaus

My old mother always used to say, my lord, that facts are like cows.
If you stare them in the face hard enough, they generally run away.
~Dorothy L. Sayers, Clouds of Witness

An Intramural Dispute (Arising from Bovine Inspection)

Most political philosophers seem to share Mrs. Bunter's view of facts. Facts about motivations, information, as well as social and institutional dynamics are often seen as pesky cows that need to be stared down so we can get on with spinning out intuitions about true normativity, natural rights or ideal justice, and exchanging those contrived stories (invoking "intuitions") at which philosophers excel. Russell Hardin long battled this absurd method of political philosophy, which renders so many of its conclusions irrelevant and useless.[1] "The worst failing of contemporary political philosophy is its frequent irrelevance to actual and plausible conditions" (Hardin 1999: 412).

Compared with the median political philosopher, Russell and I were fellow travelers. We both have insisted that any adequate view of justice or

G. Gaus (✉)
Philosophy Department, University of Arizona, Tucson, AZ, USA

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_2

morality must accommodate the facts of human life and show how notions of morality and justice facilitate, as well as regulate, myriad forms of human cooperation. And, as Russell stressed, questions of scale are critical. In his papers on "Bodo ethics" (more on these anon), he maintained that systems of moral relations that work well for small-scale closed societies might be inapplicable to large, impersonal, dynamic, societies.[2] In all this I am a Hardinite, as any reasonable political philosopher must be.

Just because we shared so many assumptions, Russell and I had grounds for fruitful debates. One of these involves the relative roles of instrumental, self-interested, rationality and social morality in explaining human cooperative social life, both small and large. Russell writes:

> In David Hume's account, our repeated resolution in the same way of an interaction in repetitive contexts may be called a convention that thereafter motivates our coordination with each other. We can commonly see conformance with such conventions as instrumentally rational and self-serving. Some of us might also eventually come to see them as morally binding. *Any such claim of morality must be a later development that comes after the instrumental motivation for following at least some of these conventions.* Even then, the moral motivation may not compel everyone; some might still be compelled primarily by their interest.

> The achievement of general social order comes prior to justice, democracy, and other systemic achievements. *It is also prior to any collection of social rules* such as Gaus addresses. These are not about priority in conceptual claims, but in causal claims. Without social order at a relatively high level, we cannot successfully establish and maintain institutions for justice, democracy, and so on. (Hardin 2013: 407–410, citation deleted, emphasis added)[3]

We need to avoid construing the disagreement between Russell and me as a pointless chicken-and-egg problem. Of course as cooperative orders increasingly secure people's interests, the tendency to comply is increased; but enhanced cooperation raises new problems (including new opportunities for cheating), which then raises new problems of coordination and cooperation that are resolved by the development of social norms and moral rules, which then further enhance the satisfaction of rational interests and allow for further fruitful coordination, and so on. In successful cooperative orders, there is a self-reinforcing relationship between advancing the basic interests of participants and normative regulation; it would be folly to suggest that one has absolute causal priority. And, of course, a quintessential

convention can evolve into a moral rule.[4] I certainly do not wish to deny that a convention might pave the way for a rule of what I have called "social morality."

The interesting dispute is, I think, how early in the development of human cooperation guidance by internalized moral rules arises, and what functions it plays when it arises. If internal guidance by what I have called "social-moral rules"[5] is at the very foundation of human cooperation, then, while of course we can accept that other modes of cooperation such as conventions play a role, we should not privilege conventions as somehow "prior" in the development of human social order. Now this notion of a way of acting being "at the foundation of human cooperation" can be parsed in a variety of ways, leading to inquiries into (1) the original locus of human cooperation—*small groups*—and the role of moral guidance in them; (2) how early moral guidance arises in *the development of agents*; and (3) how *early in human history* internalized guidance by moral rules arises. I believe internalized moral guidance arises very early in all three senses, and shall have something to say about all three, though my emphasis will be on the first. Normative guidance, shame, cheater detection and punishment are, I believe, fundamental to even the smallest cooperative orders and characterize very young moral agents and arose at the very origin of our species (Gaus 2015). Without internalized moral guidance, even small-scale cooperative orders are hopelessly inefficient and probably impossible.

I begin by examining very small groups of face-to-face cooperators: here we might think that the group is small enough, and information about the behavior of others extensive enough, so that social rules enforced simply by fear of sanctions by the group would suffice. I suggest that the data do not support this supposition: even here, the internalization of moral rules is fundamental to their cooperation and cheater suppression. I then consider Russell's charge that accounts of social cooperation based on moral rules, in which individuals act on the rules despite their interests, are stuck with invoking a variety of somewhat dubious and weak "claims of moral commitment or shared values through [to] Rawls's magical 'addition of the sense of justice and moral sentiment' to make justice work at a large scale" (Hardin 2007: 96). I shall be skeptical of this claim of Russell's, pointing to evidence in support of internalized rule compliance even in the face of high costs to personal interests, and showing that the underlying mechanisms are not especially mysterious. Lastly, I briefly turn to the fundamental issue of

how social morality functions in large-scale settings and, importantly, whether it is largely displaced by formal legal and political institutions.

Social Morality in Small-Scale Societies

Bodo

In several places, Russell depicted what he called "Bodo ethics":

> Axel Leijonhufvud ... characterizes the village society of eleventh century France in which the villager Bodo lived. We have detailed knowledge of that society from the parish records of the church of St. Germaine. Today one would say that that church is in the center of Paris, but in Bodo's time it was a rural parish distant enough from Paris that many of its inhabitants may never have seen Paris. Virtually everything Bodo consumed was produced by about eighty people, all of whom he knew well. Indeed, most of what he consumed was most likely produced by his own family. If anyone other than these eighty people touched anything he consumed, it was salt, which would have come from the ocean and would have passed through many hands on the way to St. Germaine, or it was spices, which would have traveled enormous distances and passed through even more hands. (Hardin 1999: 401–402)[6]

In different contexts, Russell focused on different features of Bodo ethics. For present purposes, the proposed underlying motivation is of interest:

> A striking feature of Bodo ethics is that it is relatively easily enforceable by the community. *An individual need not rely on self-regulation to be moral.* The knowledge that the whole community has of each individual's adherence to the local moral code allows community members to sanction miscreants. An enormous part of the debate about morality in the modern secular world is about how individuals can be motivated to act morally. That question is answered easily for Bodo's world. *The community spontaneously enforces its morality as a set of compulsory norms.* ... The exaction would typically be quick and aimed at the right person. (Hardin 2013: 412, emphasis added)

In this passage, Russell seems to advance what we might call

> *The External Moral Rules Thesis:* In a small cooperative group *G*, a system of social regulation that is seen by members of *G* as simply an *external system of moral rules* is apt to constitute an effective framework for social cooperation.

To be a bit more precise: suppose that *G* is a cooperative group between 20 and 100, in which social regulation is achieved simply through moral rules of type *E* that are generally observed (and publicly known to be) such that (*i*) members of *G* expect the typical member, Alf, to conform to *E*; (*ii*) Alf recognizes that other members of *G* expect him to conform to *E* and will usually punish Alf for infractions of *E*; yet (*iii*) Alf's only motivation for compliance with *E* is self-interest, including the fear of punishment.[7] According to the External Moral Rules Thesis such rules are generally sufficient in *G* to secure effective cooperation and social order. Note that external moral rules can be, but need not be, rules specifying a classic convention, as punishment may be necessary to secure compliance.[8]

I believe that we have strong evidence that the External Moral Rules Thesis is false. From the very beginning of human social cooperation, social order fundamentally relied on moral rules internalized by the participants. Thus, I shall dissent from Russell's claim that "In Bodo's world we do not need morality to keep us all in line because the transparency of all our actions is virtually total" (Hardin 2013: 412).

Cephu

We now possess rich ethnographic data about rules in small-scale societies. Christopher Boehm has engaged in a massive study of rules and sanctioning practices of both tribal societies and hunter-gather societies (Boehm 1999, 2012). The latter—small groups of 20–30 people—is especially interesting for us. Boehm has developed a database of over 300 hunter-gatherer societies and, of these, he has identified about half as essentially closed societies, with minimal contact with agricultural or commercial societies. These societies share much social context of Bodo's village (traditional, small, face-to-face, largely isolated)[9] except, crucially, they are not agricultural and sedentary, and much less hierarchical.

Boehm's data indicates that such small-scale societies tend to employ a hierarchy of punishments, from gossiping and criticism, ridicule, ostracism to capital punishment. He observes that, although "under the spell of Durkheim" anthropologists often depict punishment in small-scale societies as spontaneous and almost automatic, this seems mistaken. Focusing on the sanctioning of overly assertive would-be dominant individuals, Boehm holds that the typical process is considerably more political:

> First, individuals begin to grope toward a group resolution of the problem, initially by gossiping behind the deviant's back and carefully watching the reactions of others. Once consensus seems predictable, some individual still has to lead the sanctioning—unless several group members do so in concert, which can be the case with ridicule. Once in a while the deviant will be simply too intimidating—or too unpredictable—for any one person or even a small coalition to risk taking the first step. (Boehm 1999: 118)

These political dynamics are striking in Colin Turnbull's famous case of Cephu, the cheating hunter. The Pygmy hunters studied by Turnbull sometimes hunt small game with nets. The men place their nets in a long semi-circle, and women and children drive game into the nets. Cephu, having complained of consistent bad luck in hunting, decided to secretly put his nets in front of the others, so game would be first driven into his net. This worked in increasing his take but, unfortunately for him, he was observed. Turnbull continues the account as the hunters

> strode into camp with glowering faces and threw their nets on the ground outside their huts. Then they sat down, with their chins in their hands, staring into space and saying nothing. The women followed, mostly with empty baskets, but they were by no means silent. They swore at each other, they swore at their husbands, and most of all they swore at Cephu.
>
> I tried to find out what had happened, but nobody would say. Kenge, who had been sleeping, came out of our hut and joined the shouting. He was the only male who was not sitting down, and although he was young he had a powerful voice, and a colorful use of language. I heard him saying, "Cephu is an impotent old fool. No, he isn't, he is an impotent old animal—we have treated him like a man for long enough, now we should treat him like an animal. Animal!" He shouted the final epithet across at Cephu's camp, although Cephu had not yet returned.
>
> The result of Kenge's tirade was that everyone calmed down and began criticizing Cephu a little less heatedly, but on every possible score: The way he always built his camp separately, the way he had even referred to it as a separate camp, the way he mistreated his relatives, his general deceitfulness, the dirtiness of his camp, and even his own personal habits.
>
> . . .
>
> Trying not to walk too quickly, yet afraid to dawdle too deliberately, he [Cephu] made an awkward entrance. For as good an actor as Cephu it was surprising. By the time he got to the *kumamolimo* everyone was doing something to occupy himself—staring into the fire or up at the tree tops, roasting plantains, smoking, or whittling away at arrow shafts. Only Ekianga

> and Manyalibo looked impatient, but they said nothing. Cephu walked into the group, and still nobody spoke. He went up to where a youth was sitting in a chair. Usually he would have been offered a seat without his having to ask, and now he did not dare ask, and the youth continued to sit there in as nonchalant a manner as he could muster. Cephu went to another chair where Amabosu was sitting. He shook it violently when Amabosu ignored him, at which he was told, "Animals lie on the ground."
>
> . . .
>
> Cephu knew he was defeated and humiliated. Alone, his band of four or five families was too small to make an efficient hunting unit. He apologized profusely, reiterated that he really did not know he had set up his net in front of the others, and said that in any case he would hand over all the meat. This settled the matter, and accompanied by most of the group he returned to his little camp and brusquely ordered his wife to hand over the spoils. She had little chance to refuse, as hands were already reaching into her basket and under the leaves of the roof where she had hidden some liver in anticipation of just such a contingency. Even her cooking pot was emptied. Then each of the other huts was searched and all the meat taken. Cephu's family protested loudly and Cephu tried hard to cry, but this time it was forced and everyone laughed at him. He clutched his stomach and said he would die; die because he was hungry and his brothers had taken away all his food; die because he was not respected.
>
> From Cephu's camp came the sound of the old man, still trying hard to cry, moaning about his unfortunate situation, making noises that were meant to indicate hunger. From our own camp came the jeers of women, ridiculing him and imitating his moans. (Turnbull 1963: 104–108)

Note that the group decides whether a violation has occurred. Often the lead is taken by one individual, in this case Kenge, who is not necessarily the directly injured party. This helps insure that the dispute will not simply be seen a dyadic conflict.[10] Consensus then forms that a violation has occurred; note especially that while Cephu's family does not join in the punishment, neither do they resist. Because small-scale societies are a complex mix of kin and non-kin relations, and it is important that punishment does not lead to interfamily conflict. This is especially clear in cases of capital punishment, which is practiced in many hunter-gather societies.[11] In cases of capital punishment, the entire group of males, including the victim's kin, sometimes collectively kills the offender (in one noted case, the entire group, including women, participated in the execution). In many cases, a kin of the offender is selected as executioner (Boehm 1999: 81–82, 121–122, 180).[12] The critical point here is that because eruption of counter-sanctioning is

always a possibility, the rule enforced must be seen by all as legitimate, it must be agreed that a violation has occurred, and the kin of the deviant must at least passively accept, and sometimes must actively participate in, the punishment. Lethal weapons abound in hunter-gather groups, and the escalation of violence is an ever-present threat.

As Samuel Bowles and Herbert Gintis more generally stress, effective punishment depends on legitimacy: unless those to be punished and their friends and allies are convinced that the rule being enforced is legitimate and one for which community enforcement is appropriate, a punishing action taken as a means to protect social cooperation can lead to weakening it (2011: 36).[13] Experimental evidence confirms that attempts at punishment readily evoke counter-punishment when the offender does not experience guilt (Hopfensitz and Reuben 2009).

The Internalization of Moral Rules

Note that with Cephu the admission of guilt preceded the group's confiscation of his kill. Consensus on the lower levels of punishment, ridicule and mild ostracism were reached during the walk home and afterwards, and it is this less dangerous level of punishment that triggered his profuse apologies—and only after that did confiscation occur. Still, one might think, all this remains consistent with the External Moral Rules Thesis. After all, it was punishment that in the end drove Cephu to admit guilt, and Cephu was known to be something of an actor, so his profuse admissions of guilt may simply have been strategic.

The important point, though, is how costly such punishing episodes are to the group. Hunting is a highly egalitarian, cooperative, activity and shirkers, cheats, and free-riders such as Cephu pose real threats. Cephu, indeed, not only posed the threat of a cheat, but he initially resisted punishment and sought to intimidate others, arguing that he was an important person, indeed a chief (Boehm 2012: 43). Cephu, perhaps, did view the rules largely externally, and that is why he was a persistent problem.[14] Rules that were generally perceived as purely external by group members, depending solely on self-interest to motivate compliance, would be a hopelessly inefficient way of securing cooperation, inviting both opportunistic evasion and counter-punishment. The large majority must, and do, internalize the rules, which, as Boehm rightly says, involve emotional attachment to the rules and compliance with them (Boehm 2012: 113–114). Such individuals possess a virtue highly prized in many small hunter

groups—self-control.[15] In the face of temptations to cheat and dominate, they can be counted on to generally comply with the group's rules. Cephu was lacking in norm-based self-control and was a severe problem for the group: he needed watching. Those even more seriously lacking in self-control, such as repeated murderers, can be executed.[16] Overall, Boehm argues, hunter-gather societies display a high level of rule internalization and corresponding self-control.

Students of cognition have recently turned to modeling the processes that underlie norm internalization (Andrighetto et al. 2010). We know that internalization of moral rules is a normal accomplishment for humans, and occurs at a very young age. In a series of experiments conducted by Gertrud Nunner-Winkler and Beate Sodian, children between four and eight were told a story about two children, both of whom liked candy. The first child was tempted to steal the candy, but did not; the second stole the candy. Even the four-year-old subjects knew that stealing was wrong and could provide reasons why this is so. Thus they could engage in punishing violators. The difference is that the youngest children expected the child *who stole the candy to be happy* with his violation of the rule, while they (the youngest children) expected the child who *resisted temptation to be sad*. Older children reversed this; they supposed the child who stole would be sad—guilty—while the child who resisted temptation would be the happy one. Younger children apparently expect people to be happy when they get what, all things considered, they want, regardless of whether this violates a moral requirement and harms others.[17] Again, older children expected the violator to feel unhappy. Nunner-Winkler and Sodian conclude:

> children may first come to know moral rules in a purely *informational* sense, that is, they know that norms exist and why they should exist. Not until several years later, however, do they seem to treat them as personally binding obligations the intentional violation of which will be followed by negatively-charged self-evaluative emotions or genuinely empathetic concerns. (Nunner-Winkler and Sodian 1988: 1336, emphasis in original)

Very young children view moral rules as external guides, as in the External Moral Rules Thesis. They can appreciate reasons that these rules are important and even that punishment is appropriate; what they do not grasp is that the rule can function as a requirement in an agent's deliberations and can be seen as "personally binding" (Nunner-Winkler and Sodian 1988: 1324), so that the agent will feel guilt for failing to meet this

requirement even if by so doing she gets what she wants. What very young children do not grasp is that a typical moral agent cares about moral requirements and so can put aside the things that she wants and, instead, conform to the rule's requirements, and success in doing this relates to her own self-esteem. As Abraham Lincoln was said to have remarked, "when I do good, I feel good. When I do bad, I feel bad. That is my religion" (Bowles and Gintis 2011: 169).[18]

What's So Special About Hunter-Gather Societies?

I have focused on contemporary hunter-gatherer societies (with some reference to larger tribal societies), whereas Russell's "Bodo" resided in a medieval agricultural community. In trying to think about Russell's question of the "priority" of social order *v.* social morality, which is the better model? I believe the generally accepted answer is that humans evolved our technology of social cooperation within such hunter-gatherer bands, and so if our concern is some sense of priority, then it is these bands that formed the context of the evolution of human cooperation.

Just when, and why, our human ancestors became intense cooperators, is of course disputed, and so any claims we make must be highly tentative (that's the feature of facts that leads so many philosophers to try to stare them down). It is clear that humans have long been engaged in deeply cooperative hunting. Mary Stiner and her colleagues discovered distinctive differences in the bones of the carcasses of human kills between 400,000 and 200,000 years ago at Qesem Cave in Israel. Bones from carcasses from 400,000 years ago demonstrate that the human hunters employed tools to cut the meat, but the cut marks indicate the presence of a number of different cutting implements employed at different angles. Evidence from this earlier period suggests that

> meat distribution systems were less staged or canalized than those typical of Middle Paleolithic, Upper Paleolithic, and later humans. The evidence for procedural interruptions and diverse positions while cutting flesh at Qesem Cave may reflect, for example, more hands (including less experienced hands) removing meat from any given limb bone, rather than receiving shares through the butchering work of one skilled person. Several individuals may have cut pieces of meat from a bone for themselves, or the same individual may have returned to the food item many times. Either way, the feeding

> pattern from shared resources may have been highly individualized, with little or no formal apportioning of meat. (Stiner et al. 2009: 13211)

Kills from 200,000 years ago display much more uniform cut marks, indicating a single cutter, who cut and distributed the kill. A compelling hypothesis is that by this time humans were, or were well on their way to becoming, distinctly egalitarian hunters. Distribution of the kill does not seem, as in the earlier case, determined by competition among the hunters (where we can suppose the more dominant took the best, first), but by a designated cutter allocating shares of the kill (as is the case in many contemporary hunter-gather societies). To be a bit more speculative, it looks as if the socialized primate carnivores of 400,000 years ago were becoming egalitarian hunters by 200,000 years ago. It is very difficult not to conclude that egalitarian sharing of cooperative hunts had already taken root by this period. Self-control was absolutely essential to the development of such egalitarian sharing.

We have good reason to conclude (of course, tentatively, as we are always learning more about these issues) that modern, late-Pleistocene, humans lived in groups of between 25 and 150,[19] obtained a high percentage of their calories from hunting or fishing, and engaged in egalitarian meat sharing. Boehm's central thesis is that the mode of life of our common cooperative ancestors is essentially that of today's hunter-gather societies. As I have remarked, in his important study of contemporary late-Pleistocene-appropriate ("LPA") foraging societies, Boehm eliminated from consideration societies that have been heavily influenced by Western and market societies, those with some agriculture, those that trade with agricultural groups, those that rely on domesticated horses, and so on, ultimately identifying 150 (of which a third have been more minutely analyzed) contemporary forger societies whose way of life corresponds to what we know of late-Pleistocene hunter-gatherer bands (Boehm 2012: 78–82).

This assumption is certainly not uncontroversial.[20] Contemporary LPA-foraging societies exist in the Holocene era of much, much, milder climates and arguably greater ease, or at least less uncertainty, in obtaining food. In the extraordinarily harsh late-Pleistocene climate, it could well have been far less rare for groups to have faced such dire circumstances that sharing broke down, leading to the group splintering into family-sized, rather than band-sized, units, with very different evolutionary dynamics (Boehm 2012: 274ff).[21] Nevertheless, the social organization of these

societies corresponds to much of what we know about late-Pleistocene bands—they are mobile, stress sharing rather than storing meat, combine hunting with foraging and live in core bands of 20–30 persons. And some of these current LPA societies have, like late-Pleistocene bands, faced the most dire of circumstances—leading in some cases to parents eating their children (Boehm 2012: 275). At present, I believe, our best estimates of the earliest form of intense human cooperative social orders correspond to these "LPA-appropriate" hunter-gather societies, and these societies are ones in which, while punishment is a critical form of social control, it must be used carefully, its dangers mitigated by the internalization by most members of the group's rules and their self-control in the form of conscience. To put the matter bluntly: given our best current information, the evolution of social order marched hand-in-hand with the evolution of internalized social morality or, as Kitcher puts it, "normative guidance" (Kitcher 2010: Chap. 2).

To be sure, given the incredibly swift cultural evolution of the last 10,000 years[22] we cannot assume that our current social morality is anything similar to the egalitarianism of LPA societies.[23] The point, however, is that the normative competencies we find in such societies—such as norm internalization and its attendant motivation—are almost surely long-standing features of human social cooperation. So far from being odd commitments of confused, obscurantist, Kantian philosophers (Binmore 2005: vii–viii), they are universal features of cooperating groups of humans.

Normative Commitment and Sensitivity to Rules

How Muscular Is Normative Commitment?

There is, then, nothing really mysterious about a deeply cooperative species internalizing—becoming emotionally attached to—the rules that specify the terms of social cooperation, such as moral rules concerning sharing and property. This was probably fully accomplished 45,000 years ago in small groups. Having been hard on the modal political philosopher, in fairness I must observe the characteristic blind spot of many PPE-oriented philosophers, who accord an almost religious status to the manifestly false axiom that rationality concerns something like a pursuit of self-interested goals.[24] This is entailed by neither the idea of instrumental rationality nor rational choice/decision theory, and even a cursory understanding of moral psychology displays its deep implausibility. But like many widely accepted

false claims there is a genuine insight lurking here—the entirely sensible worry that such moral rule–based motivations may not be able to stand up to significant temptations to pursue one's narrow interests by defecting. "A mere norm," Russell writes, "is unlikely to override self-interest in many such contexts. Some members might be sufficiently motivated by moral commitments, but we cannot generally expect everyone to be, especially when the stakes are high" (Hardin 2013: 414). So, accepting that normal humans internalize, care about and are motivated to conform to, social morality, one may well wonder whether such merely "normative" motivation can successfully hold up to self-interest.

Social-Moral Rule Sensitivity

Cristina Bicchieri has usefully modeled this problem in terms of norm sensitivity (Bicchieri 2006: 62). Sensitivity to a norm or social rule concerns the relationship between the content/function of the rule and the moral and value commitments of a person. When a rule of social morality is strongly supported by an agent's own normative commitments, she will tend to be highly sensitive to a norm: put simply, she has many reasons for adhering to the requirements of a norm even in the face of temptations to cheat based on narrow self-interest.[25] As one's personal normative commitments and beliefs provide less support for the norm, sensitivity will decrease. A person whose only reason for compliance is fear of punishment would, on this view, tend to have a low sensitivity: he will engage in opportunistic cheating behavior when he can get away with it, or when the expectations of gain outweigh the likely punishment. Thus we can hypothesize:

> *The Justification Effect:* Alf's sensitivity to a rule of social morality tends to rise as its justification to Alf increases, where justification depends on the coherence of the rule with Alf's personal normative beliefs and convictions.

Bicchieri is clear that (what I have called) the Justification Effect varies in the population. Those with greater "reflective autonomy," she predicts, will have a stronger tendency to decrease their sensitivity to a norm as they become aware of reasons against it, while more conformist members of the group will have higher sensitivity to a rule just because, say, it has been in place for a long time, and will be less sensitive to reasons against it (Bicchieri 2016). On the other hand, as I have said, those whose sole reason to act on

the norm is the fear of punishment will have much less sensitivity to the norm and will be open to opportunistic cheating.[26]

The Justification Effect shows the importance of what I have elsewhere called "convergent normativity."[27] Many political philosophers are apt to think of the entire notion of "public reason" as a mere piece of Rawlsian jargon—and, alas, too often it is. However, it also allows us to see a fundamental feature of an effective social morality. As the rules of social morality tend toward public justification in group *G*, in the sense that overwhelmingly the members of *G* find that their personal normative beliefs and convictions support the rules, the members of *G* become more sensitive to those rules. And, so, internal motivations for compliance are stronger, and socially costly—and perhaps disruptive—punishment can be reduced. The more individuals find that the rules they live by correspond to their important personal values, moral and religious convictions, the more they are inclined to follow these rules even in cases when the rules call for significant sacrifice of their narrow interests.

This is not to deny the basic truth that, as Peter Richerson and Robert Boyd put it, "we are imperfect and often reluctant, though often very effective cooperators" (Richerson and Boyd 2008: 114). We need moral rules because we are a complex combination of selfish and cooperative creatures: the moral system, we might say, has developed on top of an earlier selfish set of motivations (Richerson and Boyd 2008; Friedman 2008: Chap. 1). Nevertheless, this moral system is real, and is a critical basis of the human cooperation. When it draws on the personal values and moral convictions of the participants, their motivational power can be channeled into social morality.

The Puzzle of Punishment

Nevertheless punishment is necessary for an effective system of social morality. Some, such as Cephu, may only be sensitive to the rules insofar as they expect punishment. More common is to have modest sensitivity, willing to abide by the rules but not at great costs, while many others have quite high sensitivity. But even they are usually concerned with self-interest, and seek ways to advance it. Boehm hypothesizes that we evolved a "flexible conscience"—able to distinguish what truly must not be done from minor violations and "exceptions" that allow us wiggle room to advance self-interest (Boehm 2012: 172–178).[28] But appealing to punishment is no quick solution to our problem. Why do people bother to punish? To be

sure, in iterated interactions based on direct reciprocity ("I'll help you if you help me, but not if you don't!"), "punishing" acts are actually elements of an optimizing strategy, and so enhance the interests of the punisher.[29] While direct reciprocity can be effective in accounting for cooperation in very small groups (dyads, triads) its capacity to sustain cooperation dramatically decreases as group size increases (Heinrich and Heinrich 2007: 51).[30] Bodo's group of 80 would seem considerably over the limit for direct reciprocity to sustain cooperation.[31] In large groups one who punishes an infraction is also following a moral rule at the cost of her own interests—she would almost surely be better off ignoring the infraction and go about her own business. So while it is certainly true that punishment is necessary to sustain a cooperative morality, it simply pushes us back to the question: why do people support morality by punishing rather than allow the infraction to pass?

Experiments have shown that we have a keen capacity to detect cheaters. And, like internalization, cheater detection is a very early human accomplishment, being manifest in three- and four-year olds (Cummins 1996a, b). And it is a capacity we effectively employ: as extensive empirical research has demonstrated, people do punish, and often at significant-to-high costs to themselves.[32] To focus on a very familiar case, in Ultimatum Games Responders will often refuse sizable stakes, and walk away with nothing rather than accept miserly offers.[33] Bicchieri has effectively argued that underlying this behavior is a concern with fairness norms (Bicchieri 2006: Chap. 3). To recall the familiar: in the United States and many other countries, one-shot Ultimatum Games result in median offers of Proposers to Responders of between 50% and 40%, with mean offers being 30–40%. Responders refuse offers of less than 20% about half the time (Bicchieri 2006: 105).[34] Play in Ultimatum Games does not importantly differ by gender or age. And, importantly for our purposes, Responder rejection rates remain high even when stakes are significantly increased. A variety of studies have shown that play in Ultimatum Games is not highly sensitive to the absolute size of the endowments being divided. In some studies, raising the stakes from, say $10 to $100 typically has no significant effect.[35] These are common results.[36] Although Responder rejection rates remains high even when playing for surprisingly high amounts, raising the stakes eventually does have the effect of decreasing rejection rates (Responders end up taking low offers rather than going away with nothing). As Steffen Andersen and his co-researchers point out, in many Ultimatum Game experiments Proposers advance very few low offers, making it difficult to judge what

Responders would do in the face of such offers. In their recent study, some treatments drastically increased the size of endowments to be divided (equivalent to 1600 hours of work in India, where the experiment took place) and they elicited many low offers by Proposers. In treatments with traditional sized stakes, the behavior of Responders was in line with normal play (though there were more low offers to be rejected); in their very high stakes treatments only 1 of 24 Responders rejected low offers (Andersen et al. 2011; Slonim and Roth 1998).

Reactive Emotions

Stakes do matter in Ultimatum Games, but it typically takes very high stakes before low offers are common and commonly accepted. Although we must accept the claim that motivations based on the internalization of social morality have their limits, in many ways such motivation is surprisingly strong in those who have been on the short-end of the unfairness stick. Why are so many Responders in Ultimatum Games so ready to deprive themselves of significant resources when there is no possibility of compensating gains through future interaction?[37] A hypothesis with strong experimental support is that reactive emotions such as anger are critical in motivating punishing behavior.[38]

Overall, I think we have good reason to accept what I shall call the *Reactive Emotion View*: Responders' rejection of low offers is partly explained in terms of Responders' emotional reaction to the offers Proposers make to *them*,[39] in particular whether the offer evokes negative emotions such as anger, irritation, or envy (Bosman et al. 2001). General theories of emotion support the anger/irritation/indignation version of this view; as Nico H. Frijda notes, anger and indignation are generally evoked by norm violation (Frijda 1996: 311). The main idea here is that, in addition to one's sensitivity to the norm, those who are on the receiving end of defection—or, by extension, those who empathize with victims—tend to get angry or irritated, and this makes them less sensitive to the costs of their punishing activities.[40]

To see this better, suppose we have a pot to be divided (X), and the Proposer's offer involves keeping n, leaving X-n to the Responder. According to the Reactive Emotions View, low offers, defined as where $X-n$ is (1) a small absolute amount and (2) n is a large proportion of X, should tend to be rejected. The personal costs of rejection are low ($X-n$ is absolutely small) but we would expect an emotional reaction because the

Proposer keeps n, which is a large percentage of X. Conversely, high offers, where $X-n$ is a sizable amount and n is a small proportion of X, should be accepted: the costs of rejection are high (the Responder would have to turn her back on a large amount) and the negative emotional reactions should be low or non-existent since the Proposer was not "greedy" (indeed, the Responder may have extra incentive to accept if her reactive emotion is joy at getting so much of X). This is the generally observed behavior.[41] But what of offers that are absolutely large, but proportionally low (i.e., when X is a very large pot, but n is a high proportion of X)? As we have seen, although Responder reactions are not highly sensitive to stakes, they do matter: rejection rates go down for very high stakes. This is consistent with the Reactive Emotions View, which depicts a trade-off rate between the costs of punishment and the negative emotions attached to being treated badly.[42] The crux of the Reactive Emotions View is that negative emotions can provide extra incentive to engage in costly punishment, not that the emotional reactions are so strong that even very large gains (say a 20% share of 1500 hours wages) will be angrily rejected. After all, we would expect that the value of monetary gains will always be increasing, but one can get only so angry: if so, at some point the value from monetary gains curve will intersect negative value of emotional reaction, leading to Responders to accept the offer, as in Fig. 1.

In Fig. 1, when the Proposer takes only a modest proportion of X, Responder does not experience negative emotions, and the monetary losses of rejection determine her decision. As Proposer claims a higher relative amount of the total pot negative emotions arises and, at point x, exceeds the monetary costs, leading the Responder to incur the monetary costs of

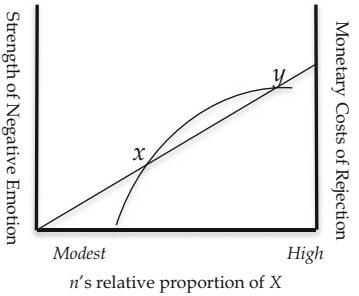


Fig. 1 Responder: monetary losses and negative emotions

rejection. However, as monetary costs of rejection continue to increase but the amount of anger at a low proportional offers does not, at point *y* the monetary costs of rejection again outweigh the negative emotions, leading Responder to accept low (proportional) offers.

If this line of reasoning is basically correct, and we also suppose that emotions are more subject to fluctuation than the costs of punishing activity (such as forgone monetary gains in the Ultimatum Game) a reasonable hypothesis is that Responders will "cool down" after a time delay. That is, we would expect Responders to accept an offer after a cool down period that they would immediately reject. The results of experiments are mixed, but I believe generally support this hypothesis. In an earlier study, a break of an hour had no effect (Bosman et al. 2001), while the more recent study of Veronika Grimm and Friederike Mengel found a marked decrease in rejection rates after only ten minutes: "While almost no low offers are accepted without delay, a large share (65–75%) of these offers gets accepted after a 10 minutes delay only" (Grimm and Mengel 2011). Grimm and Mengel also found that low offers of Proposers increase after a break; this is consistent with work on Dictator Games,[43] which indicates that Dictators whose decisions are driven by immediate affect rather than calculation make more generous offers; apparently a cool down period gives each party time to switch into calculation mode, which favors focusing on the forgone personal benefits or incurred personal costs (Schulz et al. 2014). In an experiment on the related "Power-to-Take Game" (see the following section), a more complicated pattern emerged: here both a "cooling off" and a "getting steamed up" effect seemed present. If the Proposer's actions are not too selfish from the perspective of the Responder, the Responder seems to cool off after a wait time; however, as Proposers get greedier, wait time *raises* the Responders' level of punishment (Galeotti 2013). If both cooling off and getting steamed up occur, we would expect ambiguous results from wait time experiments.

Emotions in Power-To-Take Games

A problem with measuring the role of emotions in straightforward Ultimatum Games is that Responders only have a take-it-or-leave-it choice and, as we have seen, low offers are typically uncommon. The role of emotions in Responders' behavior has been extensively studied in a "cousin" of the Ultimatum Game, the Power-to-Take Game, which allows more scope for variable emotional reaction. A Power-to-Take Game involves two players, a

Taker and a Responder; their roles are determined at random. To start, each player is given an endowment; in some treatments the players earn their endowment in a pre-game task, in others it is simply distributed by the experimenter. Suppose the endowment for each is $10. The Taker, then determines take rate—the proportion of the Responder's endowment he will take and this is announced. The Responder then has an option of destroying any amount of her endowment that she wishes, before the Taker's announced percentage is transferred from her. So if her endowment was $10, and the Taker's announced a take rate of 50%, the Taker would get $5 if the Responder destroyed none of her endowment, which would yield total payoffs of $15 for Taker and $5 for Responder. If the Responder decides to destroy half her endowment after the Taker announces his take rate, it would reduce her endowment to $5, of which the Taker would get $2.50. This game is sometimes described as an Ultimatum Game that allows variable punishment, since Responder can decide on the level at which she will deny Taker's resources.[44] But note that in this game the Responder cannot affect the Taker's endowment, but only the amount of her endowment the Taker can transfer (Reuben and van Winden 2010: 908).

In an early pioneering study by Ronald Bosman and Frans van Winden, where players earned their endowments, out of 39 subjects, only three Takers took 0, positive takings ranged from 25–100%, with a mean of 58.5%, and median 66.7%; 70% was the mode (Bosman and van Winden 2002: 153).[45] Eight Responders chose to destroy part of their endowment, and of these, *seven destroyed the entire endowment*. In a later study, Bosman, Matthias Sutter, and van Winden compared this to a game in which endowments were simply distributed at the start of play (rather than earned) (Bosman et al. 2005). Play in the no effort experiment was markedly different; Takers took an average of 32% more, and many more Responders destroyed, and more opted for intermediate destruction rates. Figure 2 summarizes the differences between the effort and no effort experiments.

	Effort	*No Effort*
Destroy Everything	7	6
Destroy Part	1	9
Destroy Nothing	31	25
Total	39	40

Fig. 2 Results in two power-to-take experiments (Bosman et al. 2005: 418)

Especially interesting is that these experiments sought to determine the extent to which emotional reactions explained behavior. Emotions were measured via self-reporting on a seven-point scale ranging from "no emotion at all" (1) to "high intensity of the emotion" (7). The emotions measured were irritation, anger, contempt, envy, jealousy, sadness, joy, happiness, shame, fear, and surprise (Bosman et al. 2005: 415).[46] The following findings are of interest to us:

> Responders who destroyed report more intense emotional reactions than those who do not.
>
> The most intense emotions of Responders who destroy in the *no effort* condition were (in order) anger, contempt, surprise and irritation.
>
> The most intense emotions of Responders who destroy in the *effort* condition were (in order) irritation, contempt, surprise and anger; the emotions tended to be more intense in this treatment.
>
> For both treatments, the intensity of these emotions is correlated with the take rate.

> With effort, the probability of destruction ... depends positively on the intensity of irritation and contempt. Without effort, the probability of destruction depends positively on the intensity of anger and contempt, and negatively on the intensity of happiness and joy. (Bosman et al. 2005: 420)

> Responders who destroy everything report more irritation than those who destroy only part. (Bosman et al. 2005: 417)

> In these studies intensity of emotional reactions is a strong predictor of Responder behavior. And importantly, anger is by no means the only relevant emotion. Especially fascinating is that contempt is always present.

In a recent study, Fabio Galeotti has shown that the predictive value of emotional reactions can be considerably lessened if the Responders' destroy options are restricted to a fixed rate (2:1) for each unit taken (Galeotti 2015). Rather than Responders deciding how much to destroy in response to a taking, they simply opt to destroy at the fixed rate or not at all. In this treatment, negative emotions remain correlated with the take rate, but have less predictive value of punishment. At low levels of punishment (for smaller takings) only contempt was of predictive value; at higher take rates (and so levels of punishment), those with higher levels of anger, irritation and contempt punished more, but this was significantly less predictive than

under variable destruction rate treatments. Fixed rate punishment thus appears to blunt the predictive effect of emotions; it especially thwarts Responders' emotionally destroying their entire endowments in response to modest takings.

Expectations and Fairness

The upshot of this line of analysis is that the mechanisms by which people uphold rules of justice and fairness at considerable costs to themselves by no means depend on a magical sense of justice; Humean limited benevolence or even simply the internalization of social rules that supports our personal normative commitments. A plausible hypothesis is that emotional reactions, especially perhaps negative ones—such as guilt by perpetrators and anger, irritation and contempt by victims—are an important foundation of upholding rules of justice among strangers.[47] However, the mere fact that in Power-to-Take Games Responders' destructive behavior is significantly, in some cases powerfully, explained by their emotional reactions does not show that emotions are related to the rules of morality and fairness. However, other data does indicate a connection. In Bicchieri's important account of social norms (roughly), a social norm is a behavioral rule *r* governing some type of behavior in a social network *S*, where most individuals in the social network prefer to conform to *r* on the conditions that (*i*) most others in *S* conform to *r* (an empirical expectation) and (*ii*) most people in *S* believe that most others in *S* ought to conform to it (a normative expectation) (Bicchieri 2006: 11). Experimental evidence involving Dictator Games indicates that when normative and empirical expectations diverge, there is a strong tendency to align behavior with the empirical expectations. An important finding in the Power-to-Take Games is that the Responders who punished very strongly tended to be (and in one study were exclusively) those who had expected lower take rates than they experienced (recall the presence of surprise) (Bosman and van Winden 2002: 156; Bosman et al. 2005: 421; Galeotti 2015: 12). This suggests that while negative emotions are well correlated with punishing behavior, this is strongly mediated by the punisher's *empirical* expectations about what others will do. However, as normative expectations have not been measured, we can only be tentative in suggesting that a norm is involved.

Thus far I have focused on Responders. Reuben and Winden studied the effect of Responders' punishment on Takers' take rate in a multi-stage Power-to-Take game (Reuben and van Winden 2010). They found that

when Responders did not destroy, the Takers who increased their take rate in the second round tended to experience *regret* after the first round—apparently regretting that they could have taken more and gotten away with it. Takers who did not experience destruction tended to increase their take rate in the second round; we might hypothesize that they were engaging in opportunistic behavior, and the absence of sanctioning encourages it. The behavior of Takers who did experience Responder destruction in the first round, however, was complex: some decreased their take rate while others did not. The key appears to be whether the Takers thought their taking was fair or unfair: those who took what they considered to be an unfair amount, to a significant degree reacted to Responders' punishment (i.e., destruction) by decreasing their takings. It is worth pointing out that in the first round these Takers apparently were willing to incur some guilt in return for high monetary gain; in the second round they may have experienced an increase in guilt, which could well have led them to lower their taking (Reuben and van Winden 2010: 918).[48] However, Responder destruction did not have the effect of lowering the take rate of those Takers who thought their takings fair. This is consistent with other studies concluding that, in addition to the anger of punishers, effective punishment requires violators to experience guilt, say in recognition that they have violated their understanding of fairness or a social norm (Hopfensitz and Reuben 2009).

I have considered experiments on Power-to-Take Games in some depth as they have focused on emotional reactions, and show that the typical fixation on anger misses a good deal of the relevant emotional reactions (and blinds us to the fascinating possibility that some reactions may be based on pride, rather than a form of moralistic aggression). We also should not make the false assumption that anger inherently leads to punishment. Experiments by Thulin and Bicchieri have shown that "moral outrage"—which is closely related to anger—underlies third-party *compensation* behavior, when norm violation has occurred. This is important: we should not suppose that negative emotions must be attached to a preference to punish violators, as opposed to compensating victims (Thulin and Bicchieri 2015).[49]

Scaling-Up Social Morality

Given that we evolved in highly cooperative small group settings, it is hardly surprising that violation of social rules is associated with significant emotional reactions. The tale of Cephu the bad hunter is about attempted

opportunistic cheating, group detection, deliberation, and the emotional storm that followed, albeit one that settled quickly once guilt had been admitted and punishment completed.[50] I have suggested that all this is far more than an ethnographer's vivid tale; the Reactive Emotions View helps explain not only the tale of Cephu, but has significant support in experimental evidence. I do not think it is much of a mystery either that we internalize moral rules and become devoted to them, or that our emotions are deeply involved in both moral judgment and action. The emotions seem especially important in inducing people to respond to defectors (Bone et al. 2014).

One still might be tempted to think this still is all about small-group settings, so it may seem that we are back to a more sophisticated version of Bodo ethics.[51] However, recall that Ultimatum and Power-to-Take Games are one-shot anonymous interactions. They are games that make sense to people habituated to non-iterated rule-based interactions with strangers. Indeed, Ultimatum Games are played fairly similarly in all large-scale market-based societies. It is when we look at very small-scale societies that we can observe marked variation. The Machiguenga (of the Amazon Basin of southeastern Peru), for example, play the game in the originally expected "selfish" way, with many lower offers that are accepted. They also play public goods games with very high rates of defection (Heinrich and Smith 2004).

The type of moral guidance that I have sketched, with internalization of group rules, concern for the legitimacy of rules, and often strong emotional reactions at being treated in ways that defy our expectations, all scale up to large-scale, anonymous interactions. It is, perhaps, precisely because in our original, and long habited, hunter-gather societies, we developed this technology of social cooperation that humans were able to so quickly and dramatically increase the scale of their societies at the beginning of the Holocene era. If the earliest societies really depended simply on rational self-interest regulated by self-interested punishment of defectors, it then *is* mysterious how humans could have left that small-scale setting for huge cooperative orders so quickly.

The answer usually given by political philosophers, is, of course, "politics and the law." In large-scale societies, it is typically held, formal institutions, not the informal framework of social morality, do the work in securing cooperation. Now of course legal and political institutions are necessary for innumerable aspects of large-scale cooperation. No sane advocate of the importance of social morality or social norms would deny that. The question

is whether these formal institutions supplant or supplement the basic framework of social rules and norms. Increasingly, I believe it is coming to be recognized that legal and political regulation without an underlying social normative framework is ineffective.[52] Gerry Mackie has pointed out that there are hundreds of critical cases around the world in which practices—among them female genital cutting, caste discrimination, child marriage—have been widely criminalized yet continue to be practiced. Laws that depart from the basic moral and social norms of a society mostly likely will be ignored, often engendering contempt for the law. As Mackie, following Iris Marion Young (2011), concludes, "Criminalization is an appropriate response to a criminal injustice, a deviation from accepted norms, its harmful consequences intended, knowingly committed by identifiable individuals, whose wrongdoing should be punished. It is not an appropriate response to a structural injustice, in compliance with accepted norms, its harmful consequences unintended byproducts, and caused by everyone and no one. The proper remedy for a harmful social norm is organized social change, not fault, blame, punishment" (Mackie 2017).

In recent years, students of social change have come to something of a consensus that effective legal regulation cannot stray too far from the underlying informal social rules.[53] One of the most striking "social experiments" based on this insight was that of Antanas Mockus, mayor of Bogotá in the late 1990s and early 2000s.[54] Mockus's aim was to harmonize legislation with social morality; he recognized that unless supported by the underlying informal moral and social framework, attempts to induce change though law would not succeed. For example, Bogotá was characterized by a very high rate of traffic fatalities in the mid-1990s, with widespread disregard for traffic regulations. Mockus distributed 350,000 "Thumbs Up/Thumbs Down" cards that drivers could display in response to dangerous driving by others, to drive home the message that such behavior was not only illegal, but violated the informal normative judgments of other drivers. Along with related programs, Bogotá witnessed a 63% decrease in traffic fatalities between 1995 and 2003. Similar programs based on harmonizing the law with informal social normative expectations led to decreases in water usage and, critically, homicides.

In lieu of an informal moral framework that coheres with the law, in a wide variety of cases (including traffic laws, which look like simply a coordination matter), we cannot expect the mass of citizens to conform unless coerced by high and effective penalties. And in the absence of such a framework, we cannot expect those occupying positions in the formal

institutions (in charge of administering those penalties) to be guided by its rules rather than taking the myriad opportunities for opportunistic enriching of themselves (Schwab and Ostrom 2008: 209–211). Institutions designed to promote cooperation can—and very often do—lead to kleptocracy (Friedman 2008: Chap. 5). Without the necessary foundation in an effective social morality, law and politics become simply additional devices by which some use power to extract from others.

Conclusion: Cows Can Be Complex

"[O]f all the differences between man and the lower animals," Darwin observes, "the moral sense or conscience is by far the most important. . . . It is the most noble of all the attributes of man. . . . Immanuel Kant exclaims, 'Duty! Wonderous thought, that worketh neither by fond insinuation, flattery, nor by any threat'" (Darwin 2004: 120). As Darwin recognized, it is the invention of morality, self-control, and conscience that allowed us to develop into one of the few eu-social species (Darwin 2004: 133). Darwin had no doubts that human morality and normative guidance was evolved, complex, and in many ways the defining feature of human social life.

Those whose work I most admire, in rightly seeking to avoid the sterility and unworldliness of so much moral and political philosophy, often turn to those models of clear-headed, empirically informed, social philosophers: Hobbes and Hume. And I freely confess that it was the hard-headed beauty of *Leviathan* that hooked me on political philosophy. For others it was the empirically rich and moderate Hume that captivated them. But while many of us deeply admire Hobbes and Hume, we must also acknowledge that their view of humans, and the ways they might solve their basic dilemmas of social life, was limited and too simple. In the last two decades, we have discovered that humans are far more complex cooperators than we thought. Recognizing the importance of social-moral rules, their internalization and enforcement, is not an appeal to the mysterious but is required by attention to the facts.

Notes

1. David Estlund (2011, 2014) explicitly accepts that the true theory of justice may well have no practical value.
2. Shaun Nichols and I have argued for this, with special reference to Bodo ethics, in Gaus and Nichols (2017).

3. For a different sort of claim that purely instrumental reasoning is in some way more basic than the notion of a rule-based social morality, see Moehler (2014).
4. Moving a bit beyond classic coordination problems, think of a "coordination" interaction such as a Stag Hunt. Even if we have achieved "hunt stag" equilibrium—which might be maintained simply by self-interest—a rule that makes it a moral requirement to hunt stag may stabilize cooperation in the face of trembling hands and other uncertainties.
5. Let us say that for a rule *R* to be a genuine social-moral rule for Betty, Betty must (*i*) recognize *R* as rule that applies to *C* circumstances; (*ii*) typically have motivating reason to conform to *R* rather than act simply on her own goals in *C* circumstances; (*iii*) her personal normative convictions endorse *R*; (*iv*) she believes that a sufficiently large subset of her group *G* conforms to *R*; (*v*) she believes that a sufficiently large subset of *G* expects her to conform to *R*. See further Gaus (2011: 163–181). In this chapter, I shall not distinguish the rules of social morality from social norms. They are not, however, equivalent; the rules of social morality are parts of practices of accountability and sustain the moral emotions of guilt, resentment and indignation; not all social norms do so.
6. See also Hardin (2013: 411ff; 2003: 98).
7. The rules are external in the sense that while agents understand them to be social guidelines that serve a purpose, their motivation to comply is simply that sanctions will be applied by others. I consider the contrast to "internalized" moral rules in more detail in section "The Internalization of Moral Rules" below.
8. See, however, the wider characterization of a convention in Bowles and Gintis (2011: 111).
9. Much depends on what is meant by a "closed" society. Marriage networks, for example, can make the group much more porous than first inspection would indicate.
10. This is one reason why "direct reciprocity" (e.g., "tit-for-tat" responses) is often a poor basis for social cooperation, engendering cycles of conflicts. See further, Boehm (2012: 60ff).
11. Boehm reports that in his database about half the hunter-gather societies are coded as having practiced capital punishment; there is strong reason to think that the number may be much higher, as central governments treat band and tribal executions as murder (2012: 84).
12. While females seldom participate in the executions, they do typically participate in the deliberation leading to execution.
13. As Bowles and Gintis point out, in large-scale societies too, anti-social punishment (counter-punishment) is real: experiments show great differences in societies to the extent to which punishment is accepted or evokes

counter-sanctioning (Ibid.). As we shall see below that in experiments in "Power-to-Take" games, Takers who were sanctioned by their partners for taking the partner's endowments but who did not see these takings as unfair, did not decrease their takings in a second round; in contrast, those who were sanctioned and did think their initial taking unfair (but hoped to get away with it) responded to sanctioning by decreasing their takings.

14. Boehm (2012: 44–45) muses that Cephu may have been something of an amoral psychopath, and so unable to internalize moral rules.
15. For a striking case, see Boehm (1999: 51–59).
16. In Boehm's database, of the societies that engaged in capital punishment, a repeat murder was the second most reported capital offense.
17. It is generally thought that young children see harm to others as violating a basic moral requirement. See Turiel et al. (1987: 174). Guilt is especially associated with violation of rules against harm and the rights of others (Prinz 2007: 77).
18. Bowles and Gintis devote much care to analyzing how internalization of social morality can be modeled (Chap. 10). As they stress, the internalization of norms is an aspect of cultural transmission that affects preferences or values. On the general phenomenon of cultural transmission, see Richerson and Boyd (2005).
19. Daniel Friedman points to 150, with much larger numbers when groups fused (Friedman 2008: 16). See also David C. Rose, who mentions 200 as the typical size of the groups in which humans evolved (Rose 2011: chp. 3). Closer examination shows that group size may be understood differently: average band size may differ from typical group size (Bowles and Gintis 2011: 95).
20. For doubts, see Richerson and Boyd (2013).
21. On the other hand, it could well have been such instability that increased the benefits of cooperation (Bowles and Gintis 2011: 93ff).
22. To what extent genes have evolved during this period is a highly controversial question. 10,000 years is far less than the 1000 generations, which is the rule-of-thumb for the evolution of major traits. But this is a highly controversial matter that is being debated (Cochran and Harpending 2009). For a rather more widely accepted view of the relation of genetic and cultural evolution, see Henrich (2016).
23. Though I have argued that it is surprisingly so (Gaus 2015).
24. Thus the common depiction of Hobbes as somehow the father of rational choice theory (even though Hobbes himself had a much more sophisticated view of human motivation). See, for example, Hartmut Kliemt, *Philosophy and Economics I: Methods and Models* (Munich: Oldenbourg, 2009), pp. 46ff.

25. On the importance of reasons, see Bicchieri and Mercier (2014).
26. Plausible models of internalization often yield polymorphic results, with a population divided between internalizers and more opportunistic types. See, for example, Andrighetto et al. (2010).
27. See, for example, Gaus (2016: chap. IV).
28. Bicchieri and her co-workers have shown how subjects exploit normative ambiguity in order to provide wiggle room to advance their interests. See Bicchieri and Chavez (2013a) and Bicchieri and Mercier (2013b).
29. As in the folk theorem, Binmore (2005: chap. 5).
30. Indirect reciprocity, or reputation, might seem to underwrite cooperation in larger groups by encouraging "boycotts" of violators, but indirect reciprocity turns out to be very sensitive to the quality of information about people. See Henrich and Henrich (2007: chap. 4), Bowles and Gintis (2011: 68–70) and Vanderschraaf (2007: 167–195).
31. In Bowles and Gintis's agent-based modeling allowing even for small rates of errors in reciprocation, groups over 10 seldom, and over 15 essentially never, evolved cooperation (2011: 64–68). Even in small group forager bands, direct reciprocity does not explain most cooperation (Boehm 2012: 179–180).
32. The experimental work on strong reciprocity and altruistic punishment is now extensive. The pioneering work was done by Ernst Fehr and his colleagues. See, for example, Fehr and Fischbacher (2005) and Fehr and Gächter (2000a, b).
33. The now famous Ultimatum Game is a single-play game between two anonymous subjects, Proposer and Responder, who have X amount of some endowment (say, money) to distribute between them. In the classic version of the game, Proposer makes the first move, and gives an offer of the form, "I will take n amount of X, leaving you with $X-n$," where n is not greater than X. If Responder accepts, each gets what Proposer offers; if Responder rejects, each receives nothing. For a recent overview see Eric van Damme et al. (2014).
34. Here some small-scale societies are outliers. See Heinrich and Smith (2004).
35. See Hoffman et al. (1996).
36. See, for example, Slonim and Roth (1998). In one study with an endowment worth three month's wages still displayed Responder rejection of lower offers (Bicchieri 2006: 114n).
37. One possible explanation—one that Russell sees as partaking of the magical—is that people may be moved by a sense of justice (Rawls 1999: chap. VIII). I do not think it is magical, and some evidence indicates that impartial concern for justice may be a motivational factor (Carlsmith et al. 2002: 284–299). Third-party punishment might be seen as based on an impartial sense of justice, and there is certainly considerable evidence for

such punishment. See also Fehr Fischbacher (2004). However, I do not think the evidence indicates this to be a critical factor, once we have factored out the reactive moral emotions, such as anger. In an interesting experiment Simon Knight sought to determine whether Responders were upholding such a sense of justice—whether "the concern is with unfair offers in general"—or were responding not to the Proposer's general status as a sharer or miser, but specifically what the Proposer did to *her*—whether the Proposer gave *her* a high or low offer. Knight finds that Responders' behavior supports the latter hypothesis—that Responder Betty's action is more strongly influenced by what has been done to *her*, so she will be apt to accept a high offer from a generally unfair Proposer or reject one from a generally fair one (Knight 2012).

38. See, for example, Hopfensitz and Reuben (2009).
39. Thus my focus at present is second-party, not third-party, punishment.
40. Another cost to which punishers appear insensitive is the number of violators; even if defection is "the norm"—there are many defectors—punishment does not generally decrease (Bone, Silva, and Raihani).
41. See, for example, Knight (2012: 7–8). As we shall see in the next section, expectations count.
42. To drastically oversimplify, The Reactive Emotion View can be modeled as claiming the decisions are based on a two-part value function. Letting $X-n$ be an offer in an Ultimatum Game, where X is the total endowment and n is the amount Proposer reserves for himself, then Responder's total value of the $X-n$ offer will be $V_{MG}-V_{RE}$, where V_{MG} is the value of the absolute monetary gain, and V_{RE} is the value based on the reactive emotions, a value arising from the negative emotions, which focus on the ratio of X to n, as mediated by expectations of what is to be expected. A Responder will accept if total value is positive, reject if it is negative.
43. In the so-called "Dictator Game" Proposer simply decides on the two shares, and that's the end of the game (not much of a game).
44. The variability of destruction is meant to uncover the relation of degree of emotional response to degree of punishment; I discuss presently a version of Power-to-Take that gives only limited punishment options which, not too surprisingly, considerably blunts the importance of emotions.
45. This is typical of takings in Power-to-Take Games; see Reuben and van Winden (2010: 912.)
46. "In both conditions, the sequence of actions was as follows. Before subjects played the one-shot PTT-game, they were randomly divided into two groups. One group was referred to as participants A (the take authorities) and the other as participants B (the responders). Subsequently, random pairs of a responder and a take authority were formed by letting take authorities draw a coded envelope from a box. The envelope contained a form on which

the endowment of both participant A and participant B was stated. The take authorities then had to fill in a take rate and put the form back in the envelope again. After the envelopes were collected, we asked the take authorities to report their emotions as well as their expectation of what the responder would do. The envelopes were brought to the matched responders who filled in the part of their endowments to be destroyed. The envelopes containing the forms were then returned to the take authorities for their information. Meanwhile, responders were asked to indicate which take rate they had expected and how intensely they had experienced several emotions after having learned about the take rate. After completing the questionnaires and collecting all envelopes, subjects were privately paid outside the laboratory by the cashier who was not present during the experiment. Experimenters were not able to see what decisions subjects made in the game and how much they earned" (Bosman et al. 2005: 415).

47. That contempt is a significant emotion in almost all experiments suggests that pride is an important explanatory character trait.
48. On the relation of guilt to interpersonal harm, see Berndsen et al. (2004).
49. It is important that Thulin and Bicchieri's target emotion appears distinctly moral; in one study emotions were measured, for example, on a 7-point scale from "Strongly Disagree" to "Strongly Agree" with statements such as "I feel angry when I learn about people suffering from unfairness" and "I think it's shameful when injustice is allowed to occur." These emotions are moral emotions, presupposing a normative content, thus in my terms they appear to function as moral rules.
50. The tale of Cephu seems to manifest both the "steaming up" and "cooling down" dynamics.
51. In some contexts, Russell intimated that the problem with all rule systems is that, because they depend on identification of a set of act-types, they cannot be usefully scaled up to regulate dynamic societies with constantly changing act-types. Nichols and I analyze this idea in Gaus and Nichols (2017).
52. I have expanded upon this point in Gaus (2018).
53. In addition to Mackie (2017), see Bicchieri (2016) and Bicchieri and Mercier (2014).
54. For a short description of this experiment, see Mockus (2012). For an in-depth treatment, see Mockus (2017).

References

Andersen, Steffen, Seda Ertaç, Uri Gneezy, Moshe Hoffman, and John A. List. 2011. Stakes Matter in Ultimatum Games. *The American Economic Review* 101: 3427–3439.

Andrighetto, Giulia, Daniel Villatoro, and Rosaria Conte. 2010. Norm Internalization in Artificial Societies. *AI Communications* 23: 325–339.

Berndsen, Mariëtte, Joop van der Pligt, Bertjan Doosje, and Antony Manstead. 2004. Guilt and Regret: The Determining Role of Interpersonal and Intrapersonal Harm. *Cognition and Emotion* 18: 55–70.

Bicchieri, Cristina. 2006. *The Grammar of Society*. Cambridge: Cambridge University Press.

———. 2016. *Norms in the Wild: How to Diagnose, Measure and Change Social Norms*. Oxford: Oxford University Press.

Bicchieri, Cristina, and Alex Chavez. 2013. Norm Manipulation, Norm Evasion: Experimental Evidence. *Economics and Philosophy* 29: 175–198.

Bicchieri, Cristina, and Hugo Mercier. 2013. Self-Serving Biases and Public Justifications in Trust Games. *Synthese* 190: 909–922.

———. 2014. Norms and Beliefs: How Change Occurs. In *The Complexity of Social Norms*, ed. Maria Xenitidou and Bruce Edmonds, 37–54. New York: Springer.

Binmore, Ken. 2005. *Natural Justice*. Oxford: Oxford University Press.

Boehm, Christopher. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press.

———. 2012. *Moral Origins: The Evolution of Virtue, Altruism and Shame*. New York: Basic Books.

Bone, Jonathan, Antonio S. Silva, and Nichola J. Raihani. 2014. Defectors, Not Norm Violators, are Punished by Third-Parties. *Biology Letters* 10. doi:10.1098/rsbl.2014.0388.

Bosman, Ronald, and Frans van Winden. 2002. Emotional Hazard in a Power-to-Take Experiment. *The Economic Journal* 112: 147–169.

Bosman, Ronald, Joep Sonnemans, and Marcel Zeelenberg. 2001. Emotions, Rejections, and Cooling Off in the Ultimatum Game. (2001) at http://hdl.handle.net/11245/1.418488

Bosman, Ronald, Matthias Sutter, and Frans van Winden. 2005. The Impact of Real Effort and Emotions in the Power-To-Take Game. *Journal of Economic Psychology* 26: 407–429.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton: Princeton University Press.

Carlsmith, Kevin M., John M. Darley, and Paul H. Robinson. 2002. Why Do We Punish?: Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology* 83: 284–299.

Cochran, Gregory, and Henry Harpending. 2009. *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution*. New York: Basic Books.

Cummins, Denise Dellarosa. 1996a. Evidence for the Innateness of Deontic Reasoning. *Mind and Language* 11: 160–190.

———. 1996b. Evidence of Deontic Reasoning in 3- and 4-Year-Olds. *Memory and Cognition* 24: 823–829.

Darwin, Charles. 2004 [1879]. *The Descent of Man*. 2nd ed. New York: Penguin.
Estlund, David. 2011. Human Nature and the Limits (if Any) of Political Philosophy. *Philosophy & Public Affairs* 39 (2011): 207–235.
———. 2014. Utopophobia. *Philosophy and Public Affairs* 42: 114–134.
Fehr, Ernst, and Urs Fischbacher. 2004. Third-Party Punishment and Social Norms. *Evolution and Human Behavior* 25: 63–87.
———. 2005. The Economics of Strong Reciprocity. In *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, ed. Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr, 151–191. Cambridge, MA: MIT Press.
Fehr, Ernst, and Simon Gächter. 2000a. Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90: 980–994.
———. 2000b. Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives* 14: 159–181.
Friedman, Daniel. 2008. *Morals and Markets: An Evolutionary Account of the Modern World*. New York: Routledge.
Frijda, Nico H. 1996. *The Emotions*. Cambridge: Cambridge University Press.
Galeotti, Fabio. 2013. An Experiment on Waiting Time and Punishing Behavior. *Economics Bulletin* 33 (2): 1383–1389.
———. 2015. Do Negative Emotions Explain Punishment in Power-To-Take Game Experiments? *Journal of Economic Psychology* 49: 1–14.
Gaus, Gerald. 2011. *The Order of Public Reason*. Cambridge: Cambridge University Press.
———. 2015. The Egalitarian Species. *Social Philosophy and Policy* 31: 1–27.
———. 2016. *Tyranny of the Ideal: Justice in a Diverse Society*. Princeton: Princeton University Press.
———. 2018. It Can't Be Rational Choice All the Way Down: Comprehensive Hobbesianism and the Origins of the Moral Order. In *Tensions in the Political Economy Project of James M. Buchanan*, ed. Peter J. Boettke, Virgil Henry Storr, and Solomon Stein. Arlington: Mercatus Center.
Gaus, Gerald, and Shaun Nichols. 2017. Moral Learning in the Open Society: The Theory and Practice of Natural Liberty. *Social Philosophy and Policy* 34.
Grimm, Veronika, and Friederike Mengel. 2011. Let Me Sleep on It: Delay Reduces Rejection Rates in Ultimatum Games. *Economics Letters* 111: 113–115.
Hardin, Russell. 1999. From Bodo Ethics to Distributive Justice. *Ethical Theory and Moral Practice* 2 (1999): 399–413.
———. 2003. *Indeterminacy and Society*. Princeton: Princeton University Press.
———. 2007. *David Hume: Moral and Political Theorist*. Oxford: Oxford University Press.
———. 2013. The Priority of Social Order. *Rationality and Society* 25: 407–421.

Henrich, Joseph. 2016. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton: Princeton University Press.

Henrich, Natalie, and Joseph Henrich. 2007. *Why Humans Cooperate*. Oxford: Oxford University Press.

Henrich, Joseph, and Natalie Smith. 2004. Comparative Evidence from Machiguenga, Mapuche, and American Populations. In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, ed. J. Henrich, R. Boyd, S. Bowles, et al., 125–167. Oxford: Oxford University Press.

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith. 1996. On Expectations and the Monetary Stakes in Ultimatum Games. *International Journal of Game Theory* 25: 289–301.

Hopfensitz, Astrid, and Ernesto Reuben. 2009. The Importance of Emotions for the Effectiveness of Social Punishment. *The Economic Journal* 119: 1534–1559.

Kitcher, Philip. 2010. *The Ethical Project*. Cambridge, MA: Harvard University Press.

Knight, Simon. 2012. Fairness or Anger in Ultimatum Game Rejections? *Journal of European Psychology Students* 3: 1–14.

Mackie, Gerry. 2017. Effective Rule of Law Requires Construction of a Social Norm of Legal Obedience. In *Cultural Agents Reloaded: The Legacy of Antanas Mockus*, ed. Carlo Tognato. Cambridge, MA: The Cultural Agents Initiative at Harvard University.

Mockus, Antanas. 2012. Building 'Citizenship Culture' in Bogotá. *Journal of International Affairs* 65: 143–146.

———. 2017. Bogotá's Capacity for Self-Transformation and Citizenship Building. In *Cultural Agents Reloaded: The Legacy of Antanas Mockus*, ed. Carlo Tognato. Cambridge, MA: The Cultural Agents Initiative at Harvard University.

Moehler, Michael. 2014. The Scope of Instrumental Morality. *Philosophical Studies* 167: 431–451.

Nunnar-Winkler, Gertrude, and Beate Sodian. 1988. Children's Understanding of Moral Emotions. *Child Development* 59: 1323–1338.

Prinz, Jesse. 2007. *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Rawls, John. 1999. *A Theory of Justice*. rev ed. Cambridge, MA: Harvard University Press.

Reuben, Ernesto, and Frans van Winden. 2010. Fairness Perceptions and Prosocial Emotions in the Power to Take. *Journal of Economic Psychology* 31: 908–922.

Richerson, Peter J., and Robert Boyd. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.

———. 2008. The Evolution of Free Enterprise Values. In *Moral Markets: The Critical Role of Values in the Economy*, ed. Paul Zak. Princeton: Princeton University Press.

———. 2013. Rethinking Paleoanthropology: A World Queerer than We Supposed. In *Evolution of Mind*, ed. Gary Hatfield and Holly Pittman, 263–302. Phildelphia: Pennsylvania Museum Conference Series.

Rose, David C. 2011. *The Moral Foundations of Economic Behavior*. New York: Oxford University Press.

Schulz, Jonathan F., Urs Fischbacher, Christian Thön, and Verena Utikal. 2014. Affect and Fairness: Dictator Games under Cognitive Load. *Journal of Economic Psychology* 41: 77–87.

Schwab, David, and Elinor Ostrom. 2008. The Vital Role of Norms and Rules in Maintaining Open Public and Private Economies. In *Moral Markets: The Critical Role of Values in the Economy*, ed. Paul Zak. Princeton: Princeton University Press.

Slonim, Robert, and Alvin E. Roth. 1998. Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic. *Econometrica* 66 (3): 569–596.

Stiner, Mary C., Ran Barkai, Avi Gopher, and James F. O'Connell. 2009. Cooperative Hunting and Meat Sharing 400–200 KYA at Qesem Cave, Israel. *Proceedings of the National Academy of Sciences of the United States of America* 106 (32): 13207–13212.

Thulin, Eric, and Cristina Bicchieri. 2015. I'm So Angry I Could Help You: Moral Outrage as a Driver of Victim Compensation. *Social Philosophy and Policy* 32 (2): 146–160.

Turiel, Elliot, Melainie Killen, and Charles C. Helwig. 1987. Morality: Its Structure, Functions and Vagaries. In *The Emergence of Morality in Young Children*, ed. Jerome Kagan and Sharon Lamb, 155–243. Chicago: Chicago University Press.

Turnbull, Colin M. 1963. *The Forest People*. New York: Simon and Schuster.

van Damme, Eric, et al. 2014. How Werner Güth's Ultimatum Game Shaped Our Understanding of Social Behavior. *Journal of Economic Behavior & Organization* 108: 292–318.

Vanderschraaf, Peter. 2007. Covenants and Reputations. *Synthese* 157: 167–195.

Young, Iris Marion. 2011. *Responsibility for Justice*. New York: Oxford University Press.

Self-Esteem

Geoffrey Brennan

When nature formed man for society she endowed him with an original desire to please and an original aversion to offend his brethren. She taught him to take pleasure in their favourable and pain in their unfavourable regard. She rendered their approbation most flattering and agreeable to him for its own sake; and their disapprobation most mortifying and most offensive.
Adam Smith [TMS p. 116].

Introduction

As far as I know, Russell Hardin never wrote specifically about esteem and the role that it plays in social and political life. But he was interested in the issue of what motivates people to act 'cooperatively' in collective action contexts where individuals' incentives seem to indicate problems for 'collective action'; and in my view, the desire for esteem plays an important explanatory role in such settings.[1] Argument in favour of that view is part of

I am grateful to Loren Lomasky, Tori McGeer, Doug MacLean, Susan Mendus and Philip Pettit for extremely helpful comments on earlier versions. Since many of those comments have been somewhat critical, these commentators deserve special exoneration!

G. Brennan (✉)
School of Philosophy, The Australian National University, Canberra, ACT, Australia

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_3

the agenda in Brennan and Pettit (2004). More broadly, that book is an attempt to develop a systematic analysis of the role that esteem plays in social and political life.

The primary focus of that earlier work was *social* esteem—the esteem that is at stake when one's actions and traits are observed by, and induce evaluative attitudes in, others. Those evaluative attitudes are termed '(positive) esteem'[2] in the case where others' attitudes are favourable and *dis*esteem when unfavourable. The esteem thereby enjoyed is taken to be an object of desire for the actor, in the spirit of the Smith quotation above. Accordingly, people will adjust their behaviour (and their character development[3]) so as to increase the positive esteem they earn from their fellows and to reduce the disesteem they receive. In other words, esteem operates as a kind of incentive—a social currency that is silently transacted as individuals undertake actions and as observers observe.

An important ambition of the earlier work was to develop analytic tools and simple models designed to capture central features of the esteem 'economy'—the social relations that govern esteem transactions, and the behavioural tendencies to which esteem incentives give rise. However, the earlier work deliberately gave only passing attention to the notion of *self*-esteem. Following the spirit of the epigraph, the earlier treatment focused on esteem as an essentially social phenomenon. On this account, as the Smith quotation suggests, it is the esteem of *others* that counts: each individual is taken to have an 'original desire to please his *brethren*' and it is the approbation '*of those brethren*' [my emphasis] that is 'most flattering and most agreeable to him'.

In the present chapter, I want to turn attention to the issue of *self*-esteem. There are two kinds of questions that are of interest in this regard:

1. To what extent do the analytical categories developed in the treatment of social esteem serve to throw interesting light on the phenomenon of self-esteem? How far can we go in treating self-esteem as just a special case of social esteem—the case, specifically, where the identity of the actor and the observer/evaluator happens to be the same? What interesting insights does such a perspective suggest?
2. When we include self-esteem in the overall picture of how esteem operates within society, how if at all does that inclusion moderate the conclusions drawn from exclusive attention to the 'social' esteem case?

In asking these questions, I am in part responding to two aspects of commentary on the earlier work (mostly in seminars and other public presentations). One aspect is that commentators often slip, apparently quite unconsciously, from using the term 'esteem' to referring to 'self-esteem' as if the two were somehow identical. A more focused and direct challenge less commonly met refers to self-esteem as the 'dog that doesn't bark' in those earlier treatments: as one critic put it, to talk of esteem without explicit attention to self-esteem is to talk of Hamlet without the Prince! Both kinds of comment suggest that somehow *self*-esteem is to be seen as the 'real game' and that social esteem is to be understood mainly as a derivative category—a view that on its face is precisely the opposite of Smith's.

One possible response to this challenge is simply to assert that self-esteem and social esteem are quite different categories, and the fact that the term 'esteem' appears in both expressions is just a linguistic accident. After all, much common use of the term self-esteem seems to emphasise the aspect of 'feeling good about oneself' without bothering to distinguish different reflexive attitudes that might be in play. In this way, self-esteem, self-confidence, self-affection and self-respect become a kind of ill-differentiated amalgam of reflexive attitudes. Moreover, self-esteem and related attitudes like self-respect often seem to play a rather different role in moral and social theory from that which social esteem plays—enrolled more in conceptualising the self than in illuminating the notion of esteem.

It is worth underlining in this connection the fact that a significant defining feature of esteem in the social setting (at least as we develop the notion[4]) is that esteem is *performance-based*: B esteems A *for a reason* (or reasons)—by virtue of A's qualities or accomplishments or actions. For B to remark that he disesteems A without being able to offer the reasons why he does so strikes a bizarre note. Perhaps B more simply just hates A, in a manner that is basically detached from reasons—but hatred is not disesteem! Esteem is precisely unlike love and hatred in this critical respect: esteem involves grounds—and quite specific grounds at that.

My instinct is to preserve, for reflexive attitudes, the same repertoire of carefully honed distinctions that we deploy in analysing people's attitudes towards others—and to insist that, even if 'self-esteem' in common parlance often does more work than is contained in the idea of social esteem turned reflexive, the notion of self-esteem ought to retain the critical feature of self-evaluation and self-assessment—by reference, however, indirect and vague, to specific performances in esteem-relevant domains.

In any event, for the purposes of the argument here, I shall take exactly that line. I shall want to treat social and self-esteem as being essentially of a piece—as involving the same kind of attitude in the observer and having the same kind of value to the performer/recipient. One implication of this strategy is that a complete treatment of overall esteem does indeed require the inclusion of self-esteem, alongside social esteem—and that a discussion of social esteem in isolation is incomplete. Just how much difference giving self-esteem its proper place will make to the analysis of overall esteem then becomes part of the explanatory agenda.

The point of departure for this investigation, then, is the stipulation that, in the case of self-esteem, the one individual plays *both* the roles relevant in the social esteem case—that is, she acts both as actor and observer. In the social esteem case, A acts in some esteem-relevant domain and is observed by B, who spontaneously forms an evaluative attitude of A's 'performance'. In the self-esteem case, A is the observer (and spontaneous appraiser) of her own performance.

To conceptualise self-esteem in this way is to treat the self as more or less continuous with other participants in the esteem 'economy'. My role as observer and evaluator of myself is *like* my role as observer and evaluator of others. Enjoying the esteem (and/or suffering from the disesteem) of others is *like* enjoying esteem from oneself (or suffering disesteem from oneself). Of course, the self may have special features in all of these roles, but the thought is that those special features can be approximated by features that might in principle be possessed by particular others. There is on this view nothing irreducible about self-esteem—or at least, if there is, that is a fact to be uncovered rather than something to be assumed *a priori*.

Consider first A in her role as observer of others. The picture is that when she observes those others operating in some esteem-relevant domain, she spontaneously forms an evaluative response. 'Great paper!' she thinks to herself. Or 'Fantastic singer!' Or 'Courageous fellow!' Or 'What a liar!' In having these thoughts, she mobilises certain values and applies certain standards.[5] The presumption is that, when the performance she witnesses happens to be her own, she will bring to bear more or less the same values and more or less the same standards as she applies in evaluating others. I say 'more or less' advisedly here, because there are well-attested elements of 'favouritism' in self-evaluation that will require more detailed treatment (specifically, in section 'The Self as Observer').

Now instead, consider the value of the esteem derived from self as against esteem derived from others. There are two aspects here. First, there is what

we might think of as a 'quality' aspect. The self may have some peculiar features as an observer—access to special knowledge or non-access to certain other knowledge—that lends the self-esteem a different quality from that deriving from 'third-persons'. Perhaps this difference between self-esteem and social esteem is rather like the difference in the esteem of someone who is an expert in the relevant domain compared with the esteem (no less intense) of someone who knows nothing about the activity.

Second, there is what we might think of as a 'value' aspect. Suppose both B and C esteem A for A's X-ing and do so equally intensely; and that B and C are equally expert in evaluating performance in the X-domain. There is nevertheless nothing to require that the given amount of 'equal-quality' esteem from B and C will be *worth* the same to A. There will certainly be some presumption to that effect if B and C are equally unknown to A and are just 'anonymous observers' on the street. Or if B and C are equally good friends of A's. But suppose that B and C differ in precisely this kind of way. We should not rule out the possibility that the strength/intimacy of A's relationship with the observer *in itself* influences the pleasure taken in any esteem received (or the pain endured in any disesteem suffered), all other things equal—and specifically the amount and quality of esteem/disesteem. And if this is so for the value A places on esteem from different observers, then there seems no reason why it could not be so for the observer A specifically. There isn't any a priori consideration suggesting that self-esteem cannot be, say, especially highly valued.

However, the idea that the esteem of certain persons is worth more than that of others, and the possibility that *self*-esteem might be especially valuable in this way, does not suggest that the value of esteem to A overall would be anything other than the sum of the value of the esteem derived from the various sources. And if this is so, it implies that social esteem and self-esteem are *substitutes*—that a higher level of social esteem can substitute to some extent for low self-esteem, and obversely a high level of self-esteem can substitute for low social esteem.[6] And that substitutability thesis is what I shall here assume.

The substitutability thesis is a claim, as we might put it, about the 'demand side' of esteem relations—about the agent's utility function in relation to esteem enjoyed. Of course, to make this claim is not to rule out the possibility of certain inter-dependencies between social esteem and self-esteem of a different kind—arising say on the supply side. However, it *is* intended to rule out two specific possibilities about the relationship between social esteem and self-esteem that seem to have some currency in the

philosophic debate. One of these is the view that self-esteem is the thing people *really* desire—and not social esteem at all: so, at best, social esteem can only play a subsidiary role. The other is the view that social esteem is an independent object of desire only in so far as the agent believes it to be 'deserved'—only in so far, that is, as the social esteem supports and endorses independently existing self-esteem. I reckon that neither of these supposed properties holds in general; and it will be the object of section 'Social Esteem and Self-Esteem' to dispose of both.

We can then proceed to examine the two basic questions that this general approach invites. First, what are the special characteristics of the self as *observer*? This question will occupy section 'The Self as Observer'. Second, what are the special characteristics of the self as a *source of esteem*? This will occupy section 'The Value of Self-Esteem'. The 'Conclusion' presents a brief summary.

At various points in the argument, I will indicate behavioural hypotheses that the conjectures suggest, and where I am aware of it, appeal to relevant evidence in relation to these hypotheses. Of course, it is quite likely that there is relevant evidence of which I am unaware—especially in the social psychological literature. But in one way, that is all to the good,[7] because *that* evidence can be treated as if it were entirely 'independent', and used either to refute or support the broader-grained conjectures. It should be clear that my aim is *not* to survey, still less to generalise upon, available empirical evidence—even though I shall help myself to relevant findings at an appropriately amateur level where I am aware of them.

Social Esteem and Self-Esteem

As earlier indicated, the main object of this section is to reject two specific accounts of the demand-side relationship between self-esteem and social esteem. However, before broaching that issue, it may help to consider a different connection, arising on the supply side.

Supply-Side Connections: The Epistemic Value of Social Esteem

In developing an attitude towards a performance and assigning esteem, there is a necessary element of judgement. I, as observer, must assess the actor's performance—note its properties and locate that performance accurately along the relevant metric of performance quality. Such a judgement need not be especially fine-grained. I can, when I go to a lieder recital, be

transported by the beauty of a voice, impressed by the intelligence and appropriateness of the interpretation, delight in the choice of repertoire—all this, without necessarily enquiring of myself whether this performance was actually superior to one by Fischer-Dieskau witnessed in 1971, or by Lisa della Casa in 1968, or Elly Ameling in 1981, all of which were comparably captivating. I just assign *this* performance to that larger 'blow-away' class and esteem the performer accordingly. But aesthetic judgements of this kind have an epistemic dimension and can be disrupted by intelligent critique by a fellow aficionado. Perhaps the tempo of the opening song was, after all, too slow. Or the vocal treatment just a bit too robust. Or the breathing in that final long phrase disappointingly obtrusive. I can have my attention drawn to these features—and either make a contrary claim or come to see the force of the critique (We singers are a tough bunch!). Equally, my judgements can be endorsed or rejected by others for whose judgement I have respect. We can wax lyrical together about just how transporting the performance was—and remind each other of the features we found especially delicious—or get down to a decent squabble.

Of course, in an environment of more or less common values, where everyone is pretty much agreed on the various dimensions of X that are relevant in forming appropriate attitudes in relation to a performance in the X-domain, having our judgements endorsed by others gives us good reason to consider them more secure. This is an idea that is formalised in the Condorcet jury theorem, and derives from the notion that, unless there is a presumption that people's judgements are on average wrong, then larger numbers of independent observations add to the epistemic authority of the predominant view. So there is good reason for me to feel that my judgement of the quality of A's X-performance is well-grounded if most others agree with my assessment. Given that the esteem others assign A reflects their judgement of A's performance in the X-domain, then the amount of esteem forthcoming operates as a proxy for their quality judgements. After all, it is a necessary feature of self-esteem that the judgements on which it is based are singular. Adding additional observers bolsters the authority of those judgements. Accordingly, my *self*-esteem will be more securely grounded if it is endorsed by *social* esteem: I can be more confident that my judgements of my own 'performance quality' are accurate if those judgements are shared by independent others, provided only that I have reason to believe that those others share my values. In this sense, social esteem and self-esteem tend to be 'complementary': self-esteem will be worth more (because more securely grounded epistemically) if it is buttressed by social esteem.

I consider this an important point; but I certainly do not think that it exhausts the relationships between social- and self-esteem. Nor of course, do I think it should be taken to suggest that, ultimately, the only thing people value is 'self-esteem'. That would be a total *non sequitor*! The evidentiary complementarity (EC) property arises on the *observer* side of the esteem relation: the claim that only self-esteem has ultimate value to the actor is a claim arising on the *actor* side of the esteem relation. This latter claim and the EC property are logically independent.

Demand-Side Factors: Three Possibilities

What is the relationship between the value of self-esteem and the value of social esteem to the performer/beneficiary? I want to lay out three possible answers to that question and try to dispose of two of them.

To distinguish these possibilities, let me set them in contrast to my preferred option.

(a) Possibility one (preferred option): Self-esteem and social esteem are basically elements of an overall aggregate 'esteem'. The self operates, in principle, just like any other observer. Of course, the self may have some unusual properties both as an observer and as a source of esteem. But on my approach, persons other than the self might, in principle, also have such properties. And indeed, differentiating between cases on the basis whether others do or do not have the relevant properties allows us to test claims about the distinctive features of self-esteem.

(b) Possibility two: Self-esteem, not social esteem, is the thing that individuals really desire. Social esteem has no independent value. But because social esteem is based on a judgement of performance level (just where in the spectrum of possibilities the actor's X-performance lies), social esteem can play the auxiliary role of providing epistemic authority to judgements relevant to self-esteem. So social esteem, or something that social esteem entails, operates as an input into self-esteem, and therefore has derivative significance. But that is its sole significance.

(c) Possibility three: Social esteem is valuable to the agent only to the extent that the agent believes it to be 'deserved'. In that sense, social esteem and self-esteem are related only in special circumstances. Self-esteem is driven by the values of the agent and by the agent's own

> assessment of her (own) performance level. If observers' values are the same as the actor's, and if observers' judgements about performance level are credible to the actor, then social esteem (or disesteem) has value (or cost) to the actor—otherwise, social esteem is irrelevant.

As already indicated in section 'Supply-Side Connections: The Epistemic Value of Social Esteem', I think (ii) gets one aspect of the relationship between self-esteem and social esteem right. Its mistake is to think that social esteem can be reduced exclusively to that role.

Possibility (iii) raises the important case, where there is heterogeneity of values—specifically, the special case where the actor's own values (call these V) diverge from those (denoted M) prevailing in the rest of society. For simplicity, consider three cases:

(a) Where V and M coincide;
(b) Where V and M are entirely independent in that they involve different but independent activities as relevant domains of evaluation. So V takes X to be a valuable activity; M takes Y to be a valuable activity (and not X). Apart from the natural limits on time and energy to devote to X and Y, which make the pursuit of X and Y competitive, X and Y are unrelated.
(c) Where V and M are directly opposed. So V holds that X is desirable; M holds that X is undesirable. Any self-esteem A enjoys by virtue of his good performance in X automatically attracts social disesteem by virtue of A's poor performance in ~X.

The three possible relationships between V and M and the three possible relationships between self- and social esteem listed in (i), (ii) and (iii) create a three-by-three matrix of possibilities that is depicted in Table 1. I shall focus on just two aspects of the comparisons across entries in the matrix: first, whether the EC property is present or not; second, on the size of the behavioural incentives in play. Such incentives can be large, medium or small and can apply to X or Y. 'Large X' is to be read as saying that the total esteem-based incentive to improve X-performance is large. 'Small X' is taken to include the possibility that the X-incentive might be negative.

The reason for focusing attention on the 'size of the incentive' in relation to activity X (and/or Y) is because that incentive (to pursue esteem) has behavioural effects, which can be used to 'test' the different models in those cases where the incentive size differs. If, for example, we want to test model

Table 1 Esteem incentives

	Esteem relations		
Value relations	*(i) Simple additivity*	*(ii) Self-esteem only*	*(iii) 'Deserved' esteem only*
(a) Identity (V = M)	EC Large X	EC Medium X	EC Large X
(b) Independence	~EC Medium X/mediumY	~EC mediumX/zeroY	~EC mediumX/zeroY
(c) Opposition	EC (negative) Small X	EC (negative) Medium X	EC (negative) Medium X

(iii) against (ii), we consider a case where the performer's values and the prevailing values in the rest of society are identical. We then consider a situation in which (e.g.) the forces of social esteem are increased—say by making A's performance more conspicuous among a larger number of people. If A devotes more effort to improving his performance in response to the social esteem thereby forthcoming, then we would reject model (ii) in favour of (i) or (iii). Or suppose we take a case where we think A's values diverge from the prevailing ones. Then making A's performance in some arena for which he has no value (or low value) more salient among a larger number of people would have no effect if (ii) and (iii) are right—but would have an effect if (i) is right.

This is what I mean by 'behavioural implications' and I am going to use these in association with relevant evidence to cast doubt on models (ii) and (iii).

Some remarks about the table. First, note that the different models of esteem relations make no difference to the EC property. Someone who, for example, thinks that only self-esteem matters will still take notice of social esteem if it has 'epistemic value'. Note too that social esteem *does* have epistemic value of a kind when V and M are directly opposed: the fact that A endures social disesteem for his good performance in X reinforces his belief that his X-performance is indeed good![8]

Second, note the differences in behavioural implications across the three models of esteem relations:

> In the case of value equality, 'deserved esteem' and 'simple additivity' will be identical—there will be a bigger behavioural effect (more X-effort) associated with social esteem in those cases than in the 'self-esteem only' case.

> In the case of independence and opposition, the 'self-esteem only' and 'deserved esteem' cases will be identical.
>
> The 'simple additivity' case will be different in two respects: A will be induced to improve Y-performance in the independence case; and A will face a net small (possibly negative) X-incentive in the opposition case.

In this light, consider the (independently interesting) case of 'pluralistic ignorance'.[9] The characteristic feature of this case is that most people believe that most other people believe that some activity X is estimable. The latter belief happens to be false; but it is false in a manner that is hard to detect, because the same forces of esteem that encourage the X behaviour discourage confession of the view that X behaviour is dis-estimable. The classic example is binge-drinking among teenagers. Most teenagers (so the argument goes) think that most of their peers think that binge-drinking is 'macho' and that to refrain from binge-drinking would be to reveal yourself as a wimp. They observe that binge-drinking is widely practised and induce from this fact that others think that binge-drinking is 'a good thing'. They comply with what is the prevailing norm. And they do not publicly question that norm, because to do so would be to reveal oneself as a wimp—the same sort of wimp that refrains from binge-drinking! In fact, however, anonymous questionnaires reveal that many of the group actually think binge-drinking is pretty disgusting. And when this fact is revealed to the group, the amount of binge-drinking goes down.

Accepting this evidence at face value, what it suggests is that social esteem is a significant behavioural influence quite apart from any epistemic value that the esteem is taken to generate. Presumably, given their actual attitudes to binge-drinking, the teenagers in the example derive some self-*dis*esteem from their drinking performances. They know that they are behaving in a manner that is 'pretty disgusting'. After all, they cannot be in any doubt about the fact that they are indeed binge-drinking.[10] But any self-disesteem is offset by what the drinkers take to be the esteem they receive from their peers. And when the peers' true values are revealed, and the social esteem drinkers are actually getting is accurately perceived, then predictably binge-drinking declines.[11]

I do not think there is anything surprising or especially controversial in the idea that social esteem is desired by individuals for its own sake (and hence that it exercises behavioural effects). To attempt to reduce everything to *self*-esteem just seems to be a mistake. But of course this is not to say that self-esteem is irrelevant or second order. The right conclusion to draw

seems to be that social esteem relates to self-esteem in potentially two ways. First, social esteem may be an *input* into self-esteem in that it may provide epistemic warrant for A's judgements about her own performance quality (and conceivably about her values, though I do not focus on this aspect here). Second, social esteem is desired in addition to self-esteem—and, as the pluralistic ignorance example indicates, not necessarily less extensively desired.

In relation to this latter aspect, it is worth bearing in mind all those cases in which it seems self-evident that being externally observed affects behaviour. So, for example, in the famous Ring of Gyges fable, Plato contends that possession of a ring that can make you invisible at will, serves to expose you to all kinds of temptations that, if visible, you would be able to resist (and indeed that might remain effectively unthinkable to you.) The important point to bear in mind in the current setting, however, is that, though the ring makes you invisible *to others*, it does not make you invisible *to yourself*! Self-esteem remains as a discipline. But not, Plato seems to have thought, a sufficiently strong one. Of course, neither I, nor Plato, need argue that self-esteem is no discipline at all. Just that social esteem plays a significant additional role.

We do not need to appeal to fables alone here. There is the interesting case of a 'hand-washing experiment' undertaken in the New York public lavatory system.[12] It was observed (via hidden cameras) that 40% of individuals who occupy the bathroom precincts on their own, washed their hands after use; whereas when the bathroom was inhabited by another person—a potential 'observer'—80% of subjects washed their hands. Now, the solitary occupant can certainly observe *herself* as a hygienic hand-washer: all the forces of self-esteem are present when she is the sole occupant. But that seems not to have the same effect as when there are others around who might observe and make judgements about the hygiene or otherwise of the actor! The obvious conclusion is that *self*-esteem is not all that is in play.

None of this is of course, enough to show that, for the demander of esteem, self-esteem and social esteem can be simply aggregated into a single 'overall esteem' measure. And actually, this is a much stronger claim than is required. My chief object here is just to cast doubt on two variants that see self-esteem as the 'major game' and social esteem as a residual category. That conclusion just does not seem to be justified empirically.

There is a further point to be made in relation to self- and social esteem, relating to the options open to individuals to increase the overall esteem

they enjoy. Short of deliberate self-deception there is probably not much you can do to increase your self-esteem—apart, that is, from improving your performance. Social esteem is different in this respect. There are a number of things you might do, *given* your performance level, to increase your social esteem. First, if your performance is in the positive esteem range, you can seek to garner greater attention—to stand in the light, as it were. [Equally, if your performance is in the negative esteem range, you can seek 'privacy/secrecy'.] Or you can seek to locate yourself in a league where your performance will appear better relative to your natural comparators than in another league where you will not shine so much. Or you can join groups that enjoy a high reputation and exploit possible economies of scale in reputational effects.

The effect of these social-esteem-augmenting strategies on incentives to perform at a yet higher level is ambiguous. On the one hand, when you secure greater attention, the esteem on offer from better performance is greater. On the other hand, attention-seeking may itself be somewhat dis-estimable; and in any event, it engages time and effort and imagination that might otherwise have been devoted to performance enhancement. Besides, when your social esteem rises to a threshold level, it might pay you to 'rest on your laurels'.[13]

The general implication of this section is that once the 'self' is treated more or less as any other observer, then social esteem and self-esteem are seen as elements of the same kind, more or less aggregating to form a measure of 'overall esteem'. This simple formulation seems broadly consistent with such evidence as I have been able to uncover on the matter, and of course has the virtues of simplicity and analytical tractability. One important implication of this formulation is that self-esteem must always be viewed against the background of the aggregate: it is social esteem *plus* self-esteem that is the behaviourally relevant parameter.

The Self as Observer

If the self is more or less the same as any other observer *conceptually,* that does not of course imply that the self as observer does not exhibit certain special features. Rather it suggests that we can explore the properties of self-esteem by thinking of properties of observers that would make an empirical difference in the social esteem case.

In that social esteem case, individuals differ as observers. Some are just more attentive as a general matter. Some are expert in the X field and

more sensitive to relevant differences in X-performance. The esteem (or disesteem) of more astute, more refined, more expert, observers will naturally count more: it is, as we might put it, of 'higher quality'. So it is natural to ask what the special characteristics, if any, of the 'self' as observer are. In addressing this question, we consider various possible such characteristics in turn.

1. **Partiality:** If we ask a number of people who are acquainted with each other to give themselves and the others a score based on certain desirable characteristics—charm, intelligence, good looks, affability, honesty, courage (etc.)—certain systematic patterns emerge. People routinely score themselves somewhat more highly than others do. In fact, there is only one group of people whose self-assessment matches what others think of them—namely, the clinically depressed. On reflection, this fact is itself somewhat depressing. The message seems to be that others find you less charming, less intelligent, less good-looking, less honest, less trustworthy than you think you are. The only consolation (if it is one) is that we seem to need this cushion of illusion to function as 'normal people'. In short, and the clinically depressed apart, we are partial to ourselves in making evaluative judgements of our own performances.[14]

There is a sense in which this self-partiality connects to our sociality. If people believe that their characteristics are more estimable than they are, then they will have esteem-seeking reason to place themselves in environments where they will be observed. If people were aware what others really made of them, then they would have (social and hence aggregate) esteem-based reasons to act more often 'in private'. And thus diminish the esteem-based incentives to behave as well as they do!

It may pay to distinguish two possible mechanisms by which this partiality might come about. First, you may have a distorted perception of your own performance—your judgement of where along the X-performance dimension you lie may suffer from sympathetic magnification. Alternatively, you might apply different standards to yourself than you apply to others—and specifically, lower standards. Perhaps you expect less of yourself. Perhaps you think that yours is a special case. Perhaps your judgement is distorted by self-affection. 'Geoff' you say to yourself. 'Not a bad chap, really. Certainly imperfect, but all those inadequacies are mere peccadilloes. And in their own way, slightly endearing.' The same kinds of distortions

that doting mothers bring to their children's performances in the primary school play, the self may bring to assessment of own performance in arenas like honesty, punctuality, consideration for others, hygiene, sense of responsibility and so on.

Or perhaps it is not so much that one's self-assessments are too generous but rather that our judgements of others are too harsh. Again, perhaps this is because we apply excessively high standards; or because we are uncharitable in our judgements of how good others' performances are—applying not so much sympathetic magnification to our own performances as rather mean-spirited de-magnification of others.

It is doubtful whether such distortions are entirely conscious: it is difficult to see how one could bring such partisan evaluations of the same quality performance to bear entirely consciously. This is why the moral life includes a measure of serious and, as far as it is possible, 'objective' self-evaluation. Still, one can be more or less hospitable to such self-examination, and more or less casual in its application. Just as some observers are less attentive or more easily distracted than are others, so some people may be especially inattentive to their own performances—or attentive in the wrong way.

Short of having any inclination towards clinical depression, however, most of us cannot be entirely unaware of the possibilities of partiality. That presumably is why for Smith (as for countless other moral theorists) it is the *impartial* spectator to whose stentorian voice moral sensibilities attend. And such partiality also suggests why social esteem might actually be more reliable than self-esteem. The neighbours may actually be a better approximation to the 'impartial spectator' than the all too cosy, benignly disposed, 'man within the breast'.[15]

2. **Epistemic Access:** We know things about ourselves that others cannot know with equal authority. Some of those things are relevant for esteem assessment. For example, to the extent that esteem attaches to motives and dispositions rather than actions, the self seems decidedly better placed to assess esteem-worthiness than any external observer—even perhaps one who knows you quite well. A knows of himself, for example, that he gives money to charity almost exclusively because he wishes to be thought generous by his peers. He calculates whether, if he refrains from making his contribution, others will find out—and what the chances are that he might get away with it, and so on. In short, A knows that he is not really a benevolent person. Others may esteem him as such, but he cannot garner any

> *self*-esteem on this front. And this consideration means that self-esteem is likely, other things being equal, to mean more to A. Social esteem is just too ill-informed to be authoritative.

There are, however, other cases where one's own judgement of 'performance' is seriously lacking—not for reasons of partiality but because of the nature of what is being observed. In the normal course of events, you cannot see your own face. And you know, from tapes replayed, that your voice sounds different to others than it sounds to you. And you are sensitive to the case of halitosis—something that, if the dental ads are right, we *have* to rely on our 'best friends' to tell us about. In such cases, you will be dependent on others to assess your 'performance'. 'How do I look?' 'How did I sound?' 'Do I have bad breath?': these are questions that you need to have answered by others before self-esteem or self-disesteem can be based on much more than mere suspicion.

In short, it shouldn't be assumed that superior epistemic authority lies either with the self or with others: this is a case-by-case matter. And of course, we can say what it is about the cases that is relevant: viz., the relative capacities of internal and external observers to make accurate judgements of performance quality. For example, consider two estimable domains—X and Z. In the X-case, esteem-worthiness is largely a matter of the agent's motivation: we give esteem not so much for the act as for the disposition that the act signifies. We might be concerned whether a person is 'truthful' in the sense that she tries to say what is true, rather than whether her beliefs are accurate: someone may say many things that happen to be false, on this reading, and still be a truthful person, because whenever she says something that is false she sincerely believes it to be true. But she knows this of herself—that she has a disposition towards truthfulness—and she can esteem herself on that basis.

In the Z-case, the relevant performance is such that the actor is necessarily a poor judge of its quality. The reasons for this might be varied: it might be that the case is like the bad breath one. Or it might be that the act of performance somehow occludes the possibility of observation. Or it might be that the performance itself is of an intrinsically relational kind such that its success depends on its bringing about some response in someone else. Being a good communicator might in certain contexts be an object of esteem: but 'good communication' is a matter of others getting your message, and requires the relevant response in the others for the communication to count as good. Being a 'good lover' might be considered

similarly. The 'good lover' or the 'good communicator' cannot assess performance unilaterally. A person knows she is good in such cases because and insofar as she is esteemed by others.

Accordingly, self-esteem will be a larger share of overall esteem in X-cases than in Z-cases: in those latter cases, social esteem predominates. But what of cases that are not especially of the X- or the Z-type—where the self has no special knowledge beyond being close to the scene of action as a matter of course? In general, there seem to be two epistemically relevant considerations; and they go in opposite directions. Social esteem has going for it the fact that the number of judges assessing your performance and locating it along the relevant spectrum is greater. You might think that a scholarly paper you have written is a 'really nice piece'—but you are only one among (we hope) potentially many readers. For reasons already mentioned, if the body of professional opinion on the paper is less enthusiastic, there is good reason to think that they are right and you are wrong—in no way different from the case in which you have an eccentric judgement of a paper *not* your own. Unless you are an especially good judge of such things (and maybe even then) the odds, other things equal, favour backing the views of the larger mass—this, provided that the average reader is more likely to make a right than a wrong assessment.

This consideration weighs in assessing particular papers. But the self has an advantage over others when it comes to making a more global assessment of your performance overall—namely, that the self is likely to have a larger sample to draw on. Sadly, there can be few readers as well acquainted with your work as you yourself. *You* have read *everything* you have written; whereas most of those others have just read a few pieces here or there. When it comes to making an assessment not just of this chapter or that, but of your general qualities as a scholar, the self has some advantages. The reason why no man is a hero to his butler (if this is indeed true) relates to the fact that the butler has a wider sample of performances to draw on. When you are an observer, sheer proximity counts! Accordingly, supposing that we are not dealing with a case where there are special problems of self-assessment (the X-type), the self has one significant advantage—that of perpetual presence. For instance, the self can know whether a failure to wash your hands after using the public lavatory is just an unfortunate lapse or whether you really are something of a dirty slob. The casual anonymous fellow-user who happens to observe your lapse cannot know this.[16]

Finally, we should note one set of cases where the value of epistemic considerations is reduced. The cases we have considered so far are ones in

which esteem (and disesteem) is assigned in larger or smaller amounts according to the level of 'performance'. Epistemic considerations bear in assessing where in the spectrum of good/bad performance a particular performance or set of performances lies. But not all esteem-relevant cases are like this. In some (like the hand-washing case), the real issue is not how assiduously you wash your hands, but whether you do so at all. Providing your observance of the prevailing norm is not so perfunctory as to not count as observance, the real issue is whether you act as required—whether you wash your hands or not, say. In cases where esteem and disesteem are based on this sort of on/off case, the epistemic demands are significantly reduced. Usually there is not much doubt about whether a subject washed her hands or not (more generally, complied with the prevailing behavioural norm or not). In such cases, most of the epistemic considerations just become second order. The one that remains is the fact of 'perpetual presence'. The self can observe 'how oft he offendeth': only very 'close' observers—butlers, spouses, parents, long-standing colleagues perhaps—compare with the self in this respect.

The aim of this section has been to explore the distinctive features of the self qua self as an observer. The object in this endeavour has been twofold: first, to isolate some of the properties of self-esteem that might distinguish it from social esteem; and second, to suggest arenas or activities where self-esteem is likely to represent the 'main game' in overall esteem stakes. Partiality is one important feature of self-esteem, though of course it is not unique to self-esteem. Some observers are likely to show partiality in assigning esteem to particular others—often because of personal connections of one kind or another (connections of affection say) that are similar in kind to the connections most normal individuals have to themselves.

The epistemic privilege issues are more complicated—precisely because they are not systematic across all cases. Several distinctions seem important. Those between:

- Esteem-relevant performances that are 'on/off' (e.g., a matter of complying with a well-defined norm) and performances that come in degrees. Epistemic considerations weigh more heavily in the latter case.
- Performances that are predominantly 'action'-defined and performances that are significantly 'motive'-defined. Social esteem is more important in the former class—self-esteem in the latter.

- Single incident cases (the quality of your appearance in tonight's opera) versus more temporally extended cases (the quality of your lifetime career). Again, self-esteem seems more significant, because more authoritative, in the latter case.

In lots of cases—instances of specific behavioural incentives based on esteem—*self*-esteem does not seem likely to play an especially predominant role. Self-esteem is most relevant for cases where the object of esteem (or disesteem) is a matter that reflects performance over a long horizon and where there are important elements of motive and disposition involved. In short, self-esteem is more important when the object of esteem is what we might refer to as a person's character. This is, perhaps, hardly an astounding conclusion. But what I think is mildly interesting is that this 'unsurprising conclusion' emerges not as a matter of direct observation but rather as a proposition derived from an examination of the special features of the self as observer. And the fact that the conclusion *is* unsurprising (if it *is* a fact) provides some support for the general picture of self-esteem that is offered here.

The Value of Self-Esteem

In the preceding section, I examined the 'quality' of self-esteem, based mainly on epistemic considerations—the authority of the judgements made that underlie esteem. In this section, I want to examine the issue of the value of esteem from different sources, other things including epistemic authority equal.

1. **The affection dimension:** The esteem of different observers may be worth different amounts to you—not just because some 'observers' are more alert, more discriminating or better placed to assess, but also because you desire *their* esteem in particular. That is, the same level of esteem with the same epistemic authority coming from two different people need not be valued to the same extent. This aspect of reality was backgrounded in the book-length analysis of social esteem because our aim was to treat esteem as a general phenomenon—not something that needed to be mediated by a catalogue of other more personal relations. But it certainly does not seem implausible that the esteem, say, of your mother matters especially to you *by virtue of* the fact that she is your mother; or the esteem of your partner *by virtue of*

> *the fact that* s/he is your partner; or the esteem of a much revered teacher, *by virtue of* his being a much revered teacher.[17]

It might seem as if this observation is similar to the discussion of 'partiality' in the previous section. But the point is quite different. It may well be that your mother is partial to you—inclined to overlook your mistakes and focus on your little successes. This means that she will, other things equal, give you more esteem than would a more objective observer. But the crucial question here is *what that esteem is worth to you*—whether the esteem of different persons, otherwise identical in 'esteem-units' is necessarily worth the same. Consider, for example, the mother who is actually niggardly in her assessments of her children—Woody Allen's mother perhaps. Her esteem may be reluctantly given, but it may genuinely matter to Mr. Allen that he gets as much of it as possible. And he will need a lot more esteem from other persons to adequately substitute for the disesteem he routinely receives from the maternal source.

In any event, the possibility that esteem must be denominated by its source for evaluative purposes is not at all implausible. And once this possibility is allowed, the question naturally arises as to the status of the 'self' in this respect. Is there any reason to think that the esteem that comes from the *self* has a special value?

At this point, issues of self-affection and self-esteem come together. I think it plausible that self-affection (an attitude that may be simply foundational and not derived from any particular considerations) may have the effect of augmenting the value that esteem from the self might possess to the self. However, it is worth noting that this augmentation effect is a two-edged sword: if I perceive my overall performance as 'below par' then the self-*dis*esteem that I suffer will also be augmented. Self-affection and self-esteem are not uniformly positively related—so the augmentation effect, if present, does not make any case for muddying the distinction between the two concepts.

2. **Testimony effects:** One source of difference that can arise in the social esteem case lies in the differential capacity of different individuals to operate as *testifiers:* the esteem of a Nobel Laureate in the academic setting, or a distinguished critic in the musical one, is likely to be more valuable simply because, if it becomes known that you are esteemed highly by that person, that tends to generate greater esteem from others. In particular, you may be able to enrol that Nobel

Laureate in writing references for you; or ensure that the distinguished critic in question is the one invited to review your performance.

And of course, esteem from such persons can increase your self-esteem, for epistemic reasons we have already examined. But here, the question is whether this 'testimonial' value is something that you can produce for yourself. And it is pretty clear that that is not the case. It is one thing for your Nobel connections to testify to your excellence and another entirely for you to do that for yourself. Self-promotion is in most contexts distinctly dis-estimable and the wise esteem-pursuer eschews it altogether.[18] As the eighteenth-century satirist, Edward Young (1968) rather nicely put it:

> *The love of praise, howe'er concealed by art,*
> *Reigns more or less and glows in every heart.*
> *The proud to gain it toils on toils endure.*
> *The Modest shun it—but to make it sure!*

In short, whatever differences in testimonial capacity there might among different possible observers, such differences provide no basis for especially valuing *self*-esteem. On the contrary, other things equal, the esteem of others will be more valued, precisely because those others are much better placed than you are to act as suppliers of favourable testimonial on your behalf.

3. **Humean multiplications:** Hume famously remarked that we value more highly the esteem of those who are themselves more highly esteemed. This may be for testimonial reasons. Or it might be because those who are highly esteemed are better judges of good performance—as evidenced in their own capacity to perform. But perhaps neither testimony nor evidentiary effects fully explain the phenomenon: it might be just a brute feature of the esteem economy.

Supposing this to be so, it offers an additional reason why social esteem might be a positive input into self-esteem. If I am highly esteemed by others, then the esteem I feel towards myself (as well as towards others) will carry more weight. If I have positive self-esteem, then the Hume effect will mean that that self-esteem will be worth more to me by virtue of the high social esteem I enjoy. But again, obversely, the social esteem I enjoy

will also increase the bite of self-disesteem should my self-evaluation be negative.

It is worth noting that when Hume talks of the esteem people enjoy, he seems to have in mind the *social* esteem they enjoy. You are unlikely to think that my esteem is especially valuable to you just because I have huge self-esteem. For these purposes at least, self-esteem does not cut much ice. If it has any influence at all, it looks to be negative: the esteem of someone who is highly esteemed by others but who evidently has low self-esteem is likely to be diminished in value. We might wonder, for example, what it is that such a person knows about himself that we do not know. In this sense, I am inclined to think that the main source of the Hume effect is evidentiary. And that Hume's point about the value of observer esteem is best understood by reference to the observer's social esteem specifically (and not overall esteem, or *a fortiori* self-esteem).

One implication for self-esteem lurking in the Humean formulation is that, if I have high self-esteem, I will be inclined to think that that self-esteem has high value. If by contrast I have low self-esteem (by which I, like most people, mean self-disesteem) then I will be inclined to think that my own attitudes to myself don't matter so much. Why care about the disesteem issuing from someone whose esteem level makes him of little account? What this means is that the value of self-esteem, aggregated across individuals in a given community, will be higher than otherwise. People with high self-esteem will value that esteem more than people with low self-esteem will suffer from their self-disesteem, ceteris paribus.

4. **Perpetual presence again:** In the previous section, mention was made of the epistemic implications of 'perpetual presence'. There, we noted that self has a larger sample of esteem-relevant episodes to draw on than the typical other, simply by virtue of the fact that the self is always around. The same fact has a more direct and simple application in relation to the self as a source of esteem. It is, I think, self-evident that the attitudes of those with whom you most commonly intersect are more significant to you than the attitudes of those with whom you intersect rarely. When, for example, you are trying to 'choose the right pond', you do so in part with an eye to the esteem you will enjoy *within* that pond. You will, when you reflect on it, be aware that others outside the pond may well have attitudes towards your performance, but these will not be salient to you in the way that the attitudes of those close around you will be. A may know that B

> thinks ill of A's honesty—but A is surrounded by persons who consider A pretty honest by most standards. A confronts B's disesteem in its full flower only on the rare occasions when A confronts B. And in that respect B's disesteem matters to A less. In short, it is the attitudes of those most immediately about you that bear most on the esteem you actually enjoy.

But of course, the self is especially privileged here. The self is your constant companion: the self's attitudes are, for better or worse, always present to you. You have for that reason especially strong motives to stand high in the eyes of self. The same might of course be said of one's spouse, or a very close professional colleague or people with whom one shares a house. These persons too have a special status—merely by virtue of being around most of the time. Of course, people can shift households and places of work and even change spouses. The self is, absent desperate extremes, impossible to shift!

The 'perpetual presence' consideration is to be distinguished from giving special weight to self-esteem as a direct object of desire. Perhaps one will have a preference for self as a source of esteem, much like a possible preference for one's mother's esteem—based purely on relational considerations. But that is a different issue—even though the points have similar effect: self-esteem looms larger as a share of total esteem as a result.

At the risk of repeating the obvious, I want to underline one conclusion of the argument in this section. I have already noted that any 'partiality' towards oneself as an esteem *source* is to be distinguished from the 'partiality in the assessment of own-performance' discussed in section 'The Self as Observer'. This distinction is not just conceptual. The 'partialities' have different effects on self-esteem incentives. Source-partiality of the kind discussed here means that self-disesteem no less than positive self-esteem will be weighed more highly. There is no necessary implication that, because the self-esteem matters more than the esteem of any other observer, the self will be more generous in assigning esteem to the self than would any external actor. Of course, knowing that this favouritism to self is likely, considerations that lend self-esteem a higher weight in overall esteem will increase the effect on overall esteem of any favouritism in play. But it is perfectly possible, even with the favouritism bias, that the self will have reason to disesteem her performance in some given arena—and if her self-esteem matters more to her, that disesteem will weigh more *heavily* with her, not more lightly!

Conclusion

The object of this chapter has been to explore the notion and some of the effects of self-esteem viewed through the lens of social esteem.

One relevant question, as I have seen it, is this: does the addition of self-esteem as an important category of esteem significantly affect the behavioural implications of esteem as a social category? Put another way, does the 'economy of esteem' look significantly different when self-esteem is given its proper (important) role in the overall picture?

My answer to this question is broadly in the negative. If the esteem one receives overall is an amalgam of self-esteem and esteem of others, the same basic structural features and the same essential behavioural implications emerge. Provided that the values that individuals use in self-assessment are the same as those they use in the assessment of others, the same general esteem incentives will be in play. And this remains true even if, as seems to be the case, individuals routinely apply rather less stringent standards to own performance than they apply to others. (i.e., they have higher self-esteem than an external assessment of their performances would warrant.)

One might respond to this answer by remarking that that is a rabbit I have already put into the hat. By framing self-esteem in the deflationary way I have, identifying the self for the purposes of the exercise as 'just another observer' and trying to isolate the special features of self-esteem by identifying the properties of self as observer/evaluator, I have shoe-horned self-esteem into a social esteem framework and that this shoe-horning loses more than it gains. I see the force of this charge in principle. In particular, I can see that viewing the self in relation to self-esteem as much like an external observer might fail to recognise the distinction between the 'internal' and the 'external' point of view—and might ultimately fail to take the constitution of self seriously. But in practice, I find the charge leaves me reasonably undisturbed. Certainly, I resist the claim that self-esteem (or even deserved esteem) is the main game in town and that social esteem is just a sort of residual category that might in congenial circumstances support the forces of self-esteem and is otherwise irrelevant. I do not think the behavioural evidence supports this view.

The other question at issue in this chapter is whether useful light can be thrown on self-esteem by examining it through the lens of social esteem. And here I am inclined to answer in the affirmative. To be sure, what one thereby analyses may be somewhat different from what passes for 'self-esteem' in popular discourse, where the distinctions between a variety of

reflexive attitudes—self-affection, self-confidence, self-recognition, self-knowledge and self-esteem—are not well-drawn. My instinct is to think that there is a case for drawing such distinctions and affixing labels appropriately—and that drawing the analogies between self-esteem and social esteem more rather than less closely is likely to be helpful in this connection.

Notes

1. See, for example, the discussion of norms and contracts in *Collective Action*, pp. 216–219.
2. The terms 'esteem', 'approbation' and 'favourable regard' are taken to be equivalent here.
3. People can be esteemed both for actions and for 'character' (or dispositions). For example, a soldier might be esteemed because he performed some heroic act—and it is the quality of the act *as such* that elicits the esteem of others. Or he might be esteemed because that act indicates that he has a courageous disposition. In this latter case, if it turned out that he had performed the act for non-courageous reasons (say, because he wanted the rewards of promotion and esteem), then the esteem forthcoming would be diminished and in the limit extinguished. For our purposes here, it won't be significant whether the object of esteem is action or disposition. But we do need a term to cover both cases: accordingly, we shall use the generic term 'performance' to cover all cases.
4. Following of course the tradition for which Smith is merely a representative spokesman. See Lovejoy (1961).
5. 'Standards' are a term of art here. For present purposes just think of the 'standard' as the performance level that separates disesteem from esteem. If the performer's performance is above standard S, the performer will be positively esteemed: if that performance is below S the performer will be disesteemed.
6. An implication is that someone who has a low level of self-esteem will have a higher demand for social esteem ceteris paribus than someone who has a high level of self-esteem.
7. This is perhaps a rather transparent attempt to make a virtue of my own ignorance—but with ignorance, what else can you do?
8. The EC property might extend beyond beliefs about performance quality to beliefs about 'value'. That is, the fact that others share your values might give you confidence in their validity. These EC domains are of course distinguishable—in row three, the opposition between V and M might confirm A's judgements of performance quality, but moderate A's confidence in her values. In that event, the esteem incentive might also be moderated, as A's

self-esteem becomes correspondingly less secure. This effect, if present, will be identical across the columns—the difference in esteem incentive levels between (i) and (ii)/(iii) will remain.

9. The classic reference is Prentice and Miller (1993) and Miller and Prentice (1994).
10. Suppose they were in such doubt. Then revealing to them that lots of their peers secretly disapproved of binge-drinking would not have any effect: they would still go on acting in the same way in ignorance of the fact that binge-drinking was what they were doing!
11. If no one thought binge-drinking was a good idea, then presumably binge-drinking would disappear. *Some* people must think binge-drinking is macho or desirable on other grounds.
12. See Munger and Harris (1989)
13. For a more extended discussion of all these matters, see Brennan and Pettit (2002, 2004).
14. There is a considerable literature on this phenomenon, by now well-attested, in relation to self-evaluation, to clinical depression and to illusions concerning degree of control over outcomes. The classic references are Taylor and Brown (1988) and Bandura (1989).
15. It is worth emphasising that it is what I take the neighbours to think of my performances, rather than what they actually *say* to me about them that is critical. Indeed, part of the source of the divergence between what others think of me and what I think of myself may well lie in a (corresponding) divergence between what others really think of me and what they say to me about such things. Norms of politeness shade into flattery: when we consider the divergence between what we say to others and what we really think about them, we have perhaps good reason to discount the things that people say to us. But such 'good reason' is equally something that it is a bit depressing to entertain. One implication is that 'social esteem' as revealed by the *signals* of esteem that are given to us by others may be as much the *cause* of self-partiality as a cure for it! I am grateful to Loren Lomasky for this point. [He may have been suggesting that his remarks to me about this paper might be a case in point!]
16. Perpetual presence has another (non-epistemic) aspect, which is taken up in the next section.
17. And, for example, not because your mother/partner is always around!
18. Of course there are certain self-promotional activities that are more or less legitimate—sending free copies of your latest book to your most influential colleagues; providing free tickets to the Opera to the most widely respected critics; and so on. But such activities are after all mediated by the belief that these people will call the shots as they see them—so that all one is subsidising is their *attention* to your performances. You are not buying (and almost certainly cannot buy) esteem as such.

References

Bandura, A. 1989. Social Cognitive Theory. In *Annals of Child Development*, *Six Theories of Child Development*, ed. R. Vasta, vol. 6, 1–60. Greenwich: JAI Press.

Brennan, G., and P. Pettit. 2002. Power Corrupts, But Can Office Ennoble? *Kyklos* 55 (2): 157–178.

———. 2004. *The Economy of Esteem: An Essay on Civil and Political Society*. Oxford: Oxford University Press.

Lovejoy, A.O. 1961. *Reflections on Human Nature*. Baltimore: Johns Hopkins University Press.

Miller, D.T., and D.A. Prentice. 1994. Collective Errors and Errors About the Collective. *Personality and Social Psychology Bulletin* 20: 541–550.

Munger, K., and S.J. Harris. 1989. Effects of an Observer on Handwashing in a Public Restroom. *Perceptual and Motor Skills* 69: 733–734.

Prentice, D.A., and D.T. Miller. 1993. Pluralistic Ignorance and Alcohol Abuse on Campus. *Journal of Personality and Social Psychology* 64: 243–256.

Taylor, S.E., and J.D. Brown. 1988. Illusion and Well-Being – A Social Psychological Perspective on Mental-Health. *Psychological Bulletin* 103 (2): 193–210.

Young, E. 1968. *The Complete Works*. Vol. 1. Hildesheim: Georg Olms.

The Freedom of the Ancients from a Humean Perspective

Bernd Lahno

Introduction

One of Russell Hardin's most forceful and influential arguments against Social Contract theory is embedded in his 'dual coordination theory' of social order. Hardin traces its origins to David Hume's theory of government, but he elaborates the argument not just in his seminal 2007 book on Hume as a moral and political theorist. While Hume never referred to or explicitly laid out the argument of the dual coordination theory, we may find it elaborated and illustrated in many places within Hardin's extended work on political order.[1] The argument is clearly Humean, but we need the sharp analytical perspective of the modern political scientist Russell Hardin to see how it evolves from Humean thought. Hence, it seems fair to credit Hardin rather than Hume with the argument's merits.

In contrast to other scholars in the Humean tradition, Hardin's interest in social order is not primarily directed at its spontaneous origins. He explicitly focuses on the range and potential of political action: that is, on intentional attempts to form communal life and its guiding rules. However, the dual coordination theory seems to suggest that all such attempts are fundamentally constrained. Forming the rules of social life almost always

B. Lahno (✉)
Universität Konstanz, Konstanz, Germany

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_4

requires the coordination of many individual efforts. So the success of any attempt to shape these rules depends heavily on the coordinative powers of a social group. And these powers in turn depend on existing conventions to facilitate coordination. As a consequence, a preference for a certain rule is not sufficient to motivate corresponding behavior. The rules that we have may not be the rules that we wish to have.

In this chapter, I will contrast this seemingly disenchanting result with the ancient ideal of liberty as conceptualized in the nineteenth century by Benjamin Constant (1819). I will argue that the dual coordination theory need not necessarily entail a pessimistic view of our potential to be politically free in the sense of the ancient. It is true that our individual power to choose the rules of social life is severely restricted. However, as David Hume's theory of artificial virtues shows, it may still be the case that we live by those rules that we wish to live by. And the fact that we live by those particular rules may actually be a consequence of the fact that they are what we wish to have.

In my argument, I will share the Humean perspective of Russell Hardin. I will take for granted a few very general and far-reaching hypotheses on the origin of social order, including the following:

> The fundamental social institutions are based on conventions which arose in (spontaneous) evolutionary processes.
>
> Conformity with a social rule is in general mutually advantageous.
>
> Individual compliance with the rules is originally motivated by interest; this individual interest may be supported by formal or informal sanctions.
>
> Government is based on an evolutionary evolved system of power and not on consent manifested in some sort of foundational agreement among the members of society.
>
> No State of Nature, no idealized Original Position can explain or justify government.
>
> There is no obligation grounded in Natural Law, nor an obligation by some fictitious Social Contract.

I adopt these hypotheses from Hume's social theory without commentating further here. Moreover, I will also follow Russell Hardin methodologically in employing a method that he masterly cultivated in his work: I will base much of my argument on an analysis of stylized game theoretical models that capture fundamental aspects of problems to do with cooperation and coordination.[2] In another regard, I will possibly transgress the Humean picture unfolded by Hardin. I shall put somewhat more weight on

the emotional underpinnings of moral institutions in Hume's work. I hope to show that such an attempt may help understanding of how and to what extent social life's conventional constraints leave room for the liberty of the ancient.

The argument proceeds as follows. After specifying the main claims of the dual coordination theory and confronting these with the ideal of the liberty of the ancients (1), I will prepare the basis for my argument by specifying forms of spontaneous order (2). Once people become aware of the benefits provided by spontaneously evolved social institutions, they will take interest in maintaining and forming these institutions. This is the awakening of politics. But dual coordination theory seems to imply that the influence which individuals may exert on fundamental social rules guiding our conduct can be neglected (3). However, as Hume argues in his theory of artificial virtues, a new motive may arise as people become mutually aware of a beneficial social practice (4). With the rise of this new motive, our constitutional interests take effect on individual compliance. Institutional practices and individual motivation become closely intertwined. The fundamental social rules and practices determine what people wish to do, but in a mature political order the rules and practices are also the result of what people wish others to do. The liberty of the ancients is feasible. I will conclude with some general remarks on the role of consent in this account of social order (5).

Dual Coordination Theory and the Liberty of the Ancients

In contrast to other Humean scholars who also see convention as the origin of social order,[3] Hardin does not focus primarily on spontaneous order. He explicitly maintains that political order is, at least in part, the result of intentional intervention. Social order is not only 'the result of human action, but not the execution of any human design' as Ferguson (1995, 119) maintained; it is also to some extent shaped by politics, in an attempt to give the social rules guiding our life a suitable and favored form (Hardin 2007, 103):

> The core of social order is not mere regularity [...] [it] is organizing institutions to produce mutually advantageous outcomes.

Dual coordination theory claims that both arenas—the one in which actions aim at forming and adjusting the rules of social interaction as well

as the one in which actions are regulated by those rules—are subject to the solution of coordination problems (Hardin 2007, 89):

> [...] to give a full explanation of the power of government, we would require a dual-coordination theory. First there is coordination among those who are the officials of the government and, second, there is coordination of the citizens in acquiescence to the government. But the latter does not require 'voluntary co-operation'—only acquiescence.

However, as this citation emphasizes, coordinating under the rules of social interaction does not require that those who are to be guided by the rules would, in some sense, consent with the rules' content. Given the rules and our expectation that others will conform, our best policy is also to conform, whether we like the rules or not.

This is, of course, a central argument against the contractarian claim that the rules of political or, more generally, social order essentially come into existence and are stabilized because of a consensual support that legitimizes these very rules. Such a support is neither necessary nor actually operative for those to be guided by the rules. Moreover, the creative powers of those politicians and government officials commissioned with the forming and maintenance of rules are also severely restricted as these individuals also operate within the range of given practices, traditions, and inherited norms which cannot be overcome by starting from scratch. In most cases, they will be capable of completing their task only if they effectively coordinate their efforts with those of others. So their endeavor will also be channeled by predefined conventional rules that guide such coordination. In essence, according to this argument, our powers to form the rules of social interaction seem severely limited on all levels of potential intervention. The shape of our communal life is essentially determined by history and tradition rather than by our interest in certain rules of conduct (Hardin 2007, 99; emphasis in original):

> [...] almost all of what we enjoy in this world is an inheritance from others who went before. Certainly most of the government that rules in any decent nation is something handed down to us, and most of us must be well served—even best served—if we keep it working well.

We inherited the social order that we inhabit. We are not given much opportunity to form this order, so the best we can do is almost always to

preserve it as it is. Compare this skeptical result with the ancient ideal of political liberty as specified by Benjamin Constant in a famous lecture of 1819:

> [The liberty of the ancients] consisted in exercising collectively, but directly, several parts of the complete sovereignty; in deliberating, in the public square, over war and peace; in forming alliances with foreign governments; in voting laws, in pronouncing judgments; in examining the accounts, the acts, the stewardship of the magistrates; in calling them to appear in front of the assembled people, in accusing, condemning or absolving them. But if this was what the ancients called liberty, they admitted as compatible with this collective freedom the complete subjection of the individual to the authority of the community.

Constant confronts this conception of 'the liberty of ancients' with what he calls 'the liberty of the moderns'. The latter actually captures the liberal understanding of today, being based on the conviction that every individual must be maximally free to lead her life without the interference of others—individually or collectively—and characterized by the recognition of individual rights and the rule of law.[4] Although conceptually opposed, the two kinds of liberty are factually compatible with each other. One can easily conceive of a society in which individuals participate equally in collectively forming the conditions of their individual and collective life, yet still remain maximally free from individual and social coercion.

However, as the last sentence of the quotation above suggests, the two liberties may equally well be in conflict with each other. There is an obvious parallel here with Isaiah Berlin's famous distinction between positive and negative liberty. As with Berlin in his famous metaphor of Sarastro's temple (1969, 145 ff).[5] Constant warns us against the latent threat of social coercion. The liberty of the ancients is a dangerous and deceptive ideal, which may well lead some people to oppress others in the name of liberty and which could easily be used to justify such attempts: This is what you as part of the body politic want, so we are acting on your will if we force you to comply. We are assisting you only in being free.

While the liberty of the ancients clearly incorporates the potential threat of terror in the name of liberty, it also displays a very attractive idea, suggesting that we consciously—if only collectively[6]—control the material and social conditions of our life and that we are not just their passive product. This, of course, is also central to the republican ideal of a free

society. However, the dual coordination theory as outlined above seems to tell us that there is little point in such an ideal. As Hardin forcefully argues, we live under those social rules that we contingently find in force. Consequently, our constitutional interests (Vanberg and Buchanan 1988) do not seem to play a significant role in forming those rules: we do not seem to have these rules because we presently wish to have them, nor because we once wished to have them, or would have wished to under certain (ideal) circumstances. People simply acquiesce to the rules they find; they do not comply with the rules for the reason that they prefer these rules to others.

In what follows, I will attempt to show that our constitutional interests may well play a significant role in the formation of social order. I will argue that there is a sense in which it might be true that:

> the rules we have are those that we wish to have,

and that

> we have these rules *because* we wish to have them.

The liberty of the ancients remains an attainable aim.

Like Hardin's argument, mine will be based on a Humean perspective on the evolution of social order. In fact, it seems to me that my argument is a supplement to Hardin's theory rather than a critique. As we will see, it leaves the central insights of the dual coordination theory untouched. In particular, it does not question its anti-contractarian implications.

Spontaneous Order

The simplest forms of behavioral regularities stem from the regularities in human nature or in the prevalent living conditions of individuals. Fishing is best at dawn, so we observe people going out to fish shortly before sunrise. Boys enter adolescence by the age of 13, so we see teenage boys turn into awkward bullies who go particularly crazy if girls are around. The two simple 2×2 normal form games in Fig. 1 may illustrate the basic structure of 'natural order'.

There is a simple and linear causal link from the situation of each individual as defined by the natural properties that he shares with others and/or by the common conditions of life to the actions chosen. A characteristic and regular behavioral pattern evolves because all individual actions

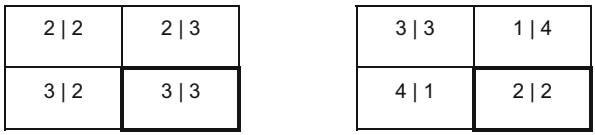


Fig. 1 Natural order

are motivated according to the same motivational pattern so defined. But any single individual action is chosen independently of every other choice. There is no social interaction in the narrow sense; actions are not motivationally related to each other. This does not mean that individuals may not be affected by what others do. The second game in Fig. 1, which is, of course, exemplifying the famous Prisoners' Dilemma (PD), illustrates this.[7] Whereas in the first game the payoff of each player is independent of the respective other's choices—the first game is best understood as a combination of two 1-person games against nature rather than as a proper 2-person game—each actor's choice in a PD has a significant (and probably drastic) impact on the outcome of the respective other. Nevertheless, it does not have a significant effect on his incentives. Defection is a dominant strategy: every actor has sufficient reason to defect, whatever the opponent chooses to do.

An interesting borderline case is the formation of a track by many individuals who make their way from A to B. They may just take the obviously shortest route, each deciding on her own judgment and essentially doing so independently of what actions she observes. But to reach B from A may require the crossing of a swamp with impassable spots which are hard to detect. In such a case individuals will be carefully testing the ground first in finding their way through the dangerous terrain. After a while, though, it will be possible to orientate oneself on the tracks found. Finally, a clear track may evolve and actors' choices will be between following the track and searching for a new way off the beaten path. That others took this path becomes a reason to do the same. Regularity thus generates an independent motivational momentum.

The paradigmatic case of a behavioral regularity that has such a feedback effect on actors is, of course, coordination. It is illustrated in Fig. 2 by the simplest 2-person coordination game, within which a given regularity makes a certain conforming behavior favorable for the individuals involved. Because others act in a certain way, it is best for me to do the same; if their behavior were different, I had better change mine, too. Although I

1 | 1	0 | 0
0 | 0	1 | 1

Fig. 2 2-Person coordination

	k or more others choose A	less than k others choose A
A	1	0
B	0	1

Fig. 3 *n*-Person coordination

may not have an interest in your aims and interests—I may have no substantial interest in what you do—I am interested in knowing what you do, because your actions set the conditions that determine how my aims are best pursued. Actions are factually interrelated here. And this produces a mutual causal relationship between individual action and social regularity. A common example of such a simple coordination problem represented by the game in Fig. 2 is the problem of choosing which side of the road to take when meeting an oncoming vehicle.

As an example of a social practice that coordinates the actions of many agents at the same time, consider a rural area in which producers offer their goods on a common, centrally located marketplace. Consumers will be looking for times when sufficiently many producers offer their goods on the market and producers, in turn, will have the best prospect of making good deals if many customers attend. For the sake of simplicity, we may assume that only two alternatives are given, A ('attend the market between 8 a.m. and noon') and B ('attend between 4 and 6 p.m.'). The decision problem of any possible market participant may (under additional simplifying assumptions) be represented by the matrix in Fig. 3. If k is sufficiently high (relative to the total number of agents) the corresponding *n*-person game has exactly two pure strategy equilibria, namely that 'all actors choose

A' or 'all actors choose B'. This suggests that a regular practice with fixed market times will evolve.

There is a significant difference in the incentive structure of the coordination problems in Figs. 2 and 3. Actors in the simple 2-person coordination game of Fig. 2 have an interest in the compliance of their partners, once a coordinative practice has evolved. They want their partners to comply if they comply themselves just as much as they themselves want to comply once they know that the partners do.[8] In the *n*-person case of Fig. 3, this relationship between individual preferences and conformity is much weaker. Although every individual agent has an interest in sufficiently many others complying, she will usually have no interest in the behavior of a particular other. I do not care when my neighbor goes to the market as long as I am certain that enough others will be there.

On the surface, such differences play no essential role in understanding coordination within the simple contexts that we discuss here. No sophisticated motivation is needed to bring about and maintain regular behavior in simple coordination problems such as those above. Interest and a reasonable expectation about others' behavior suffice. And this is true for the coordination problem in Fig. 3 just as well as for that in Fig. 2. The actual motivation we observe in real life social coordination may, nevertheless, be much more complex. People may develop additional, social motives to comply. They may, for example, have a peculiar inclination to behave as others do, which besides individual interest will give them a separate motive to conform.

The special feature in the incentive structure of our simple coordination game in Fig. 2 may account for some of the complexities that we find in the motivational structure of basic social institutions. Incentives in this game are intricately interrelated: I want to do A, because You will do A; I want You to do A, because I will do A, and I expect You to do likewise: You want to do A, because I will do A. Once individuals become aware of these correlations, two things are likely to happen. First, individuals will develop a direct interest in the conduct of others. Second, individuals will become mutually aware of the regular pattern that their actions form and perceive this pattern as a meaningful order that somehow ties together the individuals involved.

The result of this is a dramatic change in the self-conception of individuals as social beings. Individuals will realize that the actions of others contribute to a certain social order in exactly the same way as their own, and that these actions are based on individual interests and expectations

similar to their own. Consequently, others' actions are not understood as mere external constraints on one's own actions as in a game against nature. Individuals will start to *inter*act consciously with each other rather than individually maximizing their outcome in a predefined, more or less constant (social) environment.[9] Normative expectations rather than mere cognitive expectations seem psychologically natural under such conditions and informal or even formal sanctions may well evolve. Moreover, a shared understanding of the overall behavioral pattern as representing 'appropriate' or 'right' action is likely to develop. Finally, along with this, the regularity of behavior will be seen as reflecting authoritative rules of conduct. There will be a cognitive consensus about the content of these rules and a normative consensus that the rule(s) should be followed as long as compliance can be (cognitively) expected to prevail. All this will add stability to the regularity.

Compare this to Hume's description of the basic convention underlying the social institution of promising (T 522 f):[10]

> There needs but a very little practice of the world, to make us perceive all these consequences and advantages. The shortest experience of society discovers them to every mortal; and when each individual perceives the same sense of interest in all his fellows, he immediately performs his part of any contract, as being assur'd, that they will not be wanting in theirs. All of them, by concert, enter into a scheme of actions, calculated for common benefit, and agree to be true to their word; nor is there any thing requisite to form this concert or convention, but that every one have a sense of interest in the faithful fulfilling of engagements, and express that sense to other members of the society. This immediately causes that interest to operate upon them; and interest is the first obligation to the performance of promises.

Hume clearly presupposes that individuals acknowledge each other as alike in nature, being endowed with the same kind of cognitive capabilities and driven by the same kind of interests. He assumes that they share a common understanding of the beneficial practice as the outcome of a common 'scheme of actions' and know what action is required for each of the individuals involved. If all these conditions are met, Hume argues, interest alone motivates people to comply with the basic rules of the institution.

Obviously a complex social institution such as the promising institution in Hume's example copes with a coordination problem that is much more complex than any of the simple coordination games above. In the case of

promising, interaction is generally ongoing instead of 'one-shot'. Many people may be involved over the course of time although each instance of promising is a matter between two parties only. Individuals will make their choices dependent on a potentially rich source of information about their partner's behavior in the past that includes information on interactions with other actors from direct observation as well as third party testimony. The isolated single interaction—the stage game—does not show the structure of a coordination problem; here we find a severe conflict of interest between the interacting parties that constitutes a genuine dilemma. But in the ongoing interaction this changes fundamentally. As with a simple coordination game, in the resulting 'supergame' individuals face an equilibrium selection problem. As in the case of our simple toy games, there are several coordination equilibria,[11] and the objective is to coordinate actions in such a way as to select one of them.

A convention for solving such complex problems will itself constitute an intricate scheme which comprises a set of rules and roles assigning different acts to different actors in different contexts with different histories. But in principle, its character as the solution to a coordination problem remains the same. The convention defines one scheme of action among others that forms a coordination equilibrium: No individual can profit from a deviation from the scheme by one singular actor. What is more, the scheme is likely to have the stronger property of strict coordination equilibrium as in the Fig. 2 game. At any rate, it seems very plausible that any singular deviation is actually detrimental to the well-being of at least one actor (and possibly many others). So people are likely to develop a common understanding of the scheme as well as its benefits, and they will have a substantial claim on the conformance of others. Consequently, just as we believe that people are obliged to drive on the right side of the road (or on the left, as in Britain), we also think that they have an obligation to keep their word and should avoid those that do not. This contrasts with the fact that we would not expect people to care much about whether or not a particular person complies with the common market times as in the example illustrating the coordination problem in Fig. 3.

Here is a simple but profound insight: A mature conventional order goes well beyond the regular pattern of behavior resulting merely from everybody reacting rationally to what he correctly expects others to do. In a conventional order people have become aware of the mutually beneficial coordinating power of (a system of) rules. They will adjust their perception of the social world accordingly. However, although there is a shared

understanding of the 'scheme of actions', of its implications for individual decision-making and of the common benefit that mutual compliance produces, and although there might even be some normative consensus about what people should do given the conventional practice, interest remains as the ultimate force that drives individual behavior, as Hume clearly points out: I comply, because compliance is in my interest as long as (enough) others can also be expected to comply.

The Prospect of Politics

Once people are aware of the benefits made available to them by a social practice they will take interest in the cultivation of these practices. However, the benefit realized by the social practices may be controversial. Some of the practices may distribute the benefit unequally among the individuals involved, some producing less than optimal benefit. The two coordination games in Fig. 4 may illustrate these potential features.

Both games—the first game is commonly known as a 'Battle of the Sexes' (BoS), the second as a 'Hi-Lo' game—are perfect 2×2 coordination games in the sense of David Lewis. Both have two coordination equilibria (in pure strategies). All the arguments about a convention solving the equilibrium selection problem in coordination games presented above also apply to these simple games.

Assume that a practice assigns two roles, A and B ('female' and 'male'), to the players in a BoS. Assume further that the row player is always a player of role A, while a B player chooses columns. If the practice is that row and column players both choose their first alternative in the matrix, the practice will clearly favor B players. A players may well wish to change the practice to realize the other equilibrium, while B players will have an interest in maintaining the given order. However, once the practice is established, neither the wish of A for a reform nor B's interest in conserving the given practice seems to be of much relevance. As long as all As expect all Bs to stick to the practice, As can do no better than conform with the contested

1 | 2	0 | 0
0 | 0	2 | 1

1 | 1	0 | 0
0 | 0	2 | 2

Fig. 4 Contested order

practice. And as long as this is so, B individuals do not have to worry about maintenance of the practice.

The Hi-Lo game illustrates that all players involved may have an interest in a reform of a convention which, nevertheless, persists. Assume that for some contingent reason the convention is that players in a Hi-Lo situation universally choose the alternative that is listed first in the (rows or columns of the) matrix. One could, for example, imagine that the benefits were originally inverse when the practice evolved or that the practice was simply transferred from a different context in which it was (or still is) optimally beneficial for all individuals involved. If a convention is suboptimal in this way, one would expect that individuals have a common interest in bringing about change. The suboptimal convention may, nevertheless, persist. Each individual actor will optimally perform by conforming with the suboptimal practice if he has reason to expect that sufficiently many others conform with the suboptimal practice. This applies to all individuals alike and so sufficiently many may conform and all may have good reason to expect that this is actually so.

This analysis, of course, echoes a central claim of Hardin's dual coordination theory. While we may have an individual interest in changing the social order, we may even have a social consensus to affect such a change, this alone will not suffice in bringing about such a change.[12] As Hardin argues, the individual cost of initiating and organizing re-coordination will mostly be too high.[13]

The general lesson behind this argument is that the effectiveness of our constitutional interests is fundamentally constrained. A wish that society might be ordered by some preferred rule may motivate action only if actors are actually in the position—or at least believe to be in such a position—to exert a significant causal impact on the rules of interaction. This typically is not the case when interaction is guided by a convention. A singular choice between conformity with or deviation from the conventional rules simply has no significant effect on the overall practice.

The upshot is: If the fundamental rules of social life are conventional, then our primary reason to comply with these rules is not that we prefer these rules to others. As Ferguson observed, we are guided by rules that somehow evolved from our actions, but we never chose these rules. This general descriptive insight into the origin of a conventionally defined social order points to a corresponding normative one: A preference for a certain rule does not necessarily justify conforming with that rule. I may wish that all individuals including myself would act in accordance with a specific

conventional rule; however, if others do not act accordingly, it may be foolish and, in view of the consequences for all individuals involved, possibly[14] even immoral to act according to that rule's demand. And, this might be true even though all individuals agree in their evaluation of conventional rules. If all prefer conventional rule A to conventional rule B, they can still consistently and correctly expect that most will follow B. Hence, acting in accordance with A may be individually imprudent as well as immoral.[15]

However, that our preference for certain rules as a guidance in certain contexts has no direct impact on our behavior in these contexts does not mean that our corresponding constitutional interests are altogether ineffective. Hume considers the following situation in which our constitutional interests conflict with our situational inclinations (T 499):

> To the imposition then, and observance of these rules, both in general, and in every particular instance, they are at first mov'd only by a regard to interest; and this motive, on the first formation of society, is sufficiently strong and forcible. But when society has become numerous, and has increas'd to a tribe or nation, this interest is more remote; nor do men so readily perceive, that disorder and confusion follow upon every breach of these rules, as in a more narrow and contracted society.

We would prefer to go on in accordance with the established conventional rules, but given the limitations of human nature we are inclined to deviate. Can we, nevertheless, do anything to preserve the preferred practice? Here is one answer that Hume gives (T 537):

> as 'tis impossible to change or correct any thing material in our nature, the utmost we can do is to change our circumstances and situation, and render the observance of the laws of justice our nearest interest, and their violation our most remote. But this being impracticable with respect to all mankind, it can only take place with respect to a few, whom we thus immediately interest in the execution of justice.

Our preference for the traditional practice alone is not sufficient to motivate compliance. But we can change circumstances in a way that causes constitutional interests to become relevant: we enable some individuals to make choices that have a direct impact on the general compliance with rules. Government here serves the function of compensating for the imperfection of our action interests in the light of our common constitutional interests. The solution to our problem lies in intelligent institutional design.

All this is perfectly in line with the dual coordination theory. Hardin explicitly acknowledges that there is a political sphere in which individuals are directly concerned with the organization and regulation of social behavior. He points out that the installing of government (see, e.g., Hardin 2014, 86) as well as the implementation of new rules or the preservation of established rules that have become problematic (see, e.g., Hardin 2007, 89) are themselves tasks the completion of which generally requires many individual actions to be coordinated. So we implement certain measures to facilitate coordination in preferred ways. But to do so we must again employ conventions to coordinate our individual efforts in realizing the common project. As with the original coordination problems at which the whole enterprise was directed, we again face the problem that coordination might not be based on mutually preferred rules and hence might not result in mutually preferred solutions. The general insight, therefore, seems to remain in force.

Indeed, if this were the whole story it would seem to leave us with a disenchanting result: The liberty of the ancients is a fleeting ideal. Its realizations are rare and incidental. The good news is: The fundamental rules of social life are in principle mutually advantageous. The bad news is: This does not result from our individual or collective efforts to shape our social world. We are the slaves rather than the masters of life's fundamental conditions that are—more or less unconsciously—produced by the interplay of our actions.

But it is not the whole story. At least this is what I intend to argue now. And I believe that the argument can already be found in Hume's theory of moral institutions. Hume's theory elucidates how our constitutional interests may become effective even though we cannot directly control the rules that shape our social life by individual decisions.

Rules as Reasons for Action

The decisive element that Hume adds to the story told so far is his observation that a mutually beneficial social practice may fundamentally change the motivational structure of those actors who are involved in it. In short, our shared interest in beneficial and well-functioning social rules takes effect not just in the foundation of government—for instance, in establishing roles and positions, such that their holders are likely to make the public good their own. It also directly affects our willingness to comply with the rules well beyond the extent that is covered by individual interest. A new kind of

motivation emerges. The rule is internalized. It turns into an independent reason for action, into a social norm in a more narrow sense. As Hume notes in a passage from his analysis of the promising convention immediately following the quote in section 'Spontaneous Order' above (T 523):

> [. . .] interest is the first obligation to the performance of promises. Afterwards a sentiment of morals concurs with interest, and becomes a new obligation upon mankind.

In the emergence of a moral motive to comply with rules based on convention and self-interest three key factors may be distinguished, which I will briefly discuss in turn.[16]

First of all, individuals will develop a tendency to associate moral sentiments with conforming behavior because of its favorable consequences for the people involved and the public in general (T 533):

> Upon the whole, then, we are to consider this distinction betwixt justice and injustice, as having two different foundations, viz. that of self-interest, when men observe, that 'tis impossible to live in society without restraining themselves by certain rules; and that of morality, when this interest is once observe'd to be common to all mankind, and men receive a pleasure from the view of such actions as tend to the peace of society, and an uneasiness from such as are contrary to it.

Hume emphasizes here that a social practice must meet certain conditions before people start evaluating it morally. Individuals must conceive the practice as a concert of actions guided by rules; they must be aware that the practice is mutually advantageous and that all individuals involved share a constitutional interest in the maintenance of the practice. Once these conditions apply, as is likely in a mature conventional order, people start to evaluate behavior from the sole standpoint that it conforms with or deviates from the practice.

This evaluation manifests itself in moral sentiments that arise from the observation or—less agile—imagination connected with conforming or deviating behavior. The development of these sentiments is mediated by a fundamental human emotional disposition which Hume calls 'sympathy': whenever a person considers the feelings of a fellow human this is likely to strike a chord in his emotional makeup. Reflecting on another person's feelings tends to produce similar feelings in the person reflecting. These

feelings evoked by sympathy now typically undergo a transformation process in the mind of the reflecting person so that they turn into moral sentiments. Moral sentiments are essentially characterized by two features. First, they are as Hume says 'calm passions' (T 417), meaning that they are associated with a low state of arousal (which makes them easily confusable with judgment). Second, and more importantly, they are 'indirect passions' (T 276 ff), passions that are directed at an object different from their immediate cause. The original feeling was an instantaneous and direct reaction to the favorable or dismissive consequences of an individual's action. In the moral sentiment, the positive or negative character of this sentiment is preserved, but the moral sentiment is now directed toward what is perceived as the source of the good or bad consequences rather than to the consequences themselves. The characteristic objects of a moral sentiment are found in the act that brought about the consequences and, ultimately, the person who is perceived to be the genuine origin of this act. Moral sentiments, thus, represent an evaluation of an act and, ultimately, the person acting.

By sympathy and the emergence of moral sentiments, the feelings of those directly affected by conforming or deviating behavior spread among those observing or merely imagining this behavior. Thus, a shared beneficial practice yields a common scheme of emotional reaction: individuals will generally appreciate conforming behavior and disapprove of deviating behavior, even though they might not be directly affected by the consequences of the behavior being evaluated.

The role of the two other factors in the emergence of the new motive is primarily to generalize and unify the emotional reactions with conforming or deviating behavior and, thus, to foster the common practice of moral evaluation.

More precisely, the second factor is based on the human mind's propensity to perpetuate a regularity experienced in the past. One may comprehend this propensity which Hume refers to as the impact of 'custom or habit' (see, e.g., E 43 f. as the result of Pavlovian conditioning in the sphere of thought. The human mind tends to stick to the paths imprinted by past experience. The relevant consequence here is that humans are rule-following animals. They are inclined to follow the lead of a rule without considering every singular case (T 499):

> The general rule reaches beyond those instances, from which it arose.

This tendency applies also to the processes underlying moral approbation (T 585):

> Where a character is, in every respect, fitted to be beneficial to society, the imagination passes easily from the cause to the effect, without considering that there are still some circumstances wanting to render the cause a compleat one. General rules create a species of probability, which sometimes influences the judgement, and always the imagination.

Hence, a person observing conforming behavior will be inclined to display a positive emotional reaction, although she might not in fact observe corresponding positive consequences. But, of course, initially such positive consequences created the bond between observing conforming behavior and feeling approbation. The emergence of a moral sentiment out of the immediate positive emotions caused by the good consequences of a specific kind of action is associated with a fundamental change in the role of good consequences. They do not just lose their function as the intentional object of an observer's emotional reaction. Through the influence of 'custom', they also lose their role as its cause and necessary condition. In the end, the emotional reaction is not a reaction to the good consequences observed, but rather becomes a direct reaction to the act, which is experienced as 'right'. Or, to put it in terms of the subject's perspective: the good consequences lose their role as reasons for moral approbation. A certain kind of act is approved of for the sole reason that it conforms with the guiding rules of the practice.[17]

Finally, the third factor further fosters the unification of moral approbation across the individual actors involved in a social practice. It is directed at the discord in the approbation of others' conduct because of individual involvement: (T 585):

> [. . .] 'tis impossible we cou'd ever converse together on any reasonable terms, were each of us to consider characters and persons, only as they appear from his peculiar point of view.

The cure for this obstacle is based in a natural human inclination to overcome such dissonances.[18] And the means that people exploit is adopting the perspective of an uninvolved observer (T 581 f):

> [...] we fix on some steady and general points of view; and always, in our thoughts, place ourselves in them, whatever may be our present situation.

Thus, by imagination and reflection we are able to correct (T 583) our sentiments to some extent and, more rigorously, our language as well as judgments (T 585; E 227) such that some sort of consensus about the moral quality of actions is attainable. The consensus about the advantages of the common practice is complemented with a consensual practice of moral approbation.

Hume's overall argument is based on various hypotheses about the psychological nature of human beings. Whether or not these hypotheses are correct is an empirical question that cannot be answered by abstract argument based on game theoretical models. However, modern psychological theories such as the theories of learning or cognitive dissonance avoidance seem to affirm Hume's claims.

If Hume is right, then, the important consequence is: A new motive fostering compliance with a conventionally grounded social practice may arise once individuals become aware that the practice is mutually beneficial as a whole independently of the particular consequences of individual acts. This new motive is based on moral sentiments and constitutes, as Hume says (T 523), 'a new obligation upon mankind'. It is effective, transcending what is warranted by individual interest and the particular consequences of individual acts. We conceive the guiding rules of the practice as defining what is 'right'. We take actions to enforce the rules and, therefore, we deliberately comply with the rules even though we know that this might be disadvantageous in the individual case.

With this new motive, our constitutional interest to live in a world guided by certain rules translates into action interests which, to some extent, guide our choices under the respective rules. However, as Hume also notes, the new motive may well not be strong enough to overcome effectively the force of self-interest and the misguidance by a shortsighted or partisan perspective. So Hardin's argument seems to retain its force: the costs of making choices in conformity with what we would wish to be the rule may still be too high even if we feel morally inclined or morally obliged to comply. It is in this context that Hume introduces the idea of government as a remedy to the weaknesses of human nature (T 537). The new motive is, therefore, complemented by the possibility of institutional design.

So, we found that there are three ways in which our constitutional interests may become effective in forming the social order:

> The 'new motive' grounded in our shared constitutional interests may directly motivate us to comply with the rules preferred beyond what is warranted by individual interest.
>
> It may also motivate the imposition of positive as well as negative sanctions on those complying or deviating even though sanctions might be costly for those sanctioning.
>
> Constitutional interests may directly motivate investing effort in institutional design and reform. And again, the 'new motive' may reinforce this motivation by enhancing the individual willingness to bear the costs of reform.

Each of these forces toward a realization of our constitutional interests in living under certain rules is of limited effectiveness, as Hume confirms. But in combination and by mutually reinforcing each other they may, as Hume also points out, form and stabilize a mutually beneficial practice.

This does not mean that we live by the rules that we chose. We cannot choose the rules individually. And even if there were some collective choice procedure to determine the preferred rules (which is not the case), a collective choice using this procedure would not on its own warrant conformity with the rule. That a rule actually guides a social practice is constituted by a multitude of individual decisions to comply. No social rule is put into practice by mere decree. Moreover, Hardin's argument that we simply inherited most of the practices and norms that guide our social life remains valid. Yet, the practices and rules that constitute social order together with our individual preferences and aspirations are not as one-directionally related as one may think. It is true that our preferences and choices are formed by the social practices we find. However, these social practices are also formed by our choices, some of which are motivated by our approval of these practices and some of which are intentionally directed at them. There is a bi-directional causal relationship between the social order and our individual preferences. Our social practices and our constitutional interests in the rules that constitute the practices co-evolve. This leaves us with the prospect that:

- the rules we have may finally coincide with the rules that we wish to have; and that
- we may ultimately have these rules (at least in part), because these are the rules that we wish to have.

The liberty of the ancients is, therefore, a consistent and approachable ideal.

Conclusion

My argument was motivated by an impression that the dual coordination theory leaves no room for political freedom in the sense of the liberty of the ancients. I then argued that Hume's theory of moral institutions actually preserves the possibility of the liberty of the ancient. In a mature moral institution, people approve of its fundamental, mutually beneficial rules. If there is a consensus to this effect, a new motive arises: individuals will want to comply with the rules even at a cost and they will also be willing to invest some effort in making others comply. Thus, the fundamental rules of a moral institution acquire normative force. We act in certain ways because this is what the rule says and we demand corresponding conduct from others for the same reason. Hence, our consensual (albeit still individual) wish to live under the guidance of these rules explains to some extent that the rules are actually valid: not only do we comply with them but we also understand and accept their normative demand.

Does this mean that the dual coordination theory is false? I do not think it does. As I have pointed out here, my argument is perfectly consistent with the central claims of the dual coordination theory. Social order is constituted by social practices and rules that guide and constrain individual decisions in solving fundamental social coordination problems. Any attempt to form the practices and their guiding rules poses itself a coordination problem. So the rules of society cannot be chosen deliberately; they are the result of a complex social process nobody can fully and directly control. I did not deny any of these claims. However, with Hume I maintain that the process in which the rules evolve also forms our motivation. The new motive that arises according to Hume's theory may then re-affect the process in which it evolved. The picture that emerges is dynamic: A social practice sets the constraints and forms individual motivation to act. The choices so defined form and possibly reform the practice. Dual coordination theory emphasizes that all choices in this process are constrained by the underlying practice, which is not chosen according to constitutional interests. I agree. But I add: Although the practice is not deliberately chosen, it is formed by a process in which the mutual approval of the guiding rules plays a significant motivational role; we all consciously and deliberately participate in maintaining and forming a mutually beneficial practice because we mutually approve of the practice.

This may sound to some like reaffirming a version of the Social Contract theory. But this is a misconception. I argued that mutual approval and, thus,

individual consent and social consensus do play a *causal* role in the emergence of the new motive. We identified the new motive as causally effective in the forming and maintenance of a moral institution. Consent and consensus may, therefore, contribute to an explanation of the respective social practice. They are actually a necessary part of a sufficient explanation of the practice inasmuch as the new motive is a necessary element of such an explanation. However, it is not true that consent on its own or some form of (tacit) agreement have significant motivational or justificatory power, as a contractarian would typically affirm.[19] The new motive that stems from the approbation of compliance to a preferred rule is conditional; it will become motivationally effective only if individuals expect sufficiently many others to share their approbation. Otherwise, they cannot expect to be able to coordinate in the desired ways. But the fact that the motivational force of the new motive depends on the perception of mutual consent does not imply that consent has any independent motivational force on its own. The argument from dual coordination theory retains its full force: We comply because, given what we expect others to do, compliance is the optimal way to further our aims (now including those aims defined by the new motive) not because there is consent. The consent does not define our aims (as it does in a contract) and, thus, cannot acquire motivational force. And a similar argument shows that consent does not have justificatory force either (cf. the corresponding argument in section 'The Prospect of Politics').

The new motive that evolves according to Hume's theory is an individual motive stimulating individual choices. But it is not defined by our individual interests. Compliance with the rule that forms the core of the new motive must certainly be mutually advantageous. But the rule is not the result of reconciling individual interests as in a contract. It is part of a shared perspective onto the world that evolves in a common social practice.

Notes

1. See for example Hardin (1995), 28 ff. or Hardin (2014). The theory is also clearly present, albeit not explicitly stated, in other earlier work such as Hardin (1999).
2. See Hardin (1988, 1989, 1999, 2004, 2007).
3. As, for example, Robert Sugden (2009), Ken Binmore (2005) or Antony de Jasay (2010); for the general discussion see also Lahno and Brennan (2013/14).

4. Note that both kinds of liberty apply to individuals. The liberty of the ancients is 'collective' in the sense that it focuses on participatory rights in forming the collective body and its guiding rules. But these rights are still individual rights. The liberty of the ancients is about what individuals can do, albeit what they can do as part of the body politic.
5. Arguably, Constant's distinction is much more suited to accentuate this problem than its modern counterpart. As Berlin notes, the threat of unjust coercion in the name of liberty is not essentially tied to positive conceptions of liberty (1969, 134). Coercion may conceivably be justified in similar ways by reference to the ideal of negative liberty. In contrast, the liberty of the moderns as defined by Constant is conceptually inconsistent with the violation of liberal rights.
6. Many thinkers outside the liberal tradition thought that, once the world was civilized, we can control the determinants of our life only collectively and they held variants of the liberty of the ancients, therefore, to be the only achievable forms of liberty within a mature human society. See, for example Rousseau's concept of 'liberté civile' (Contrat Social 1.8, see Rousseau 1987, 159) or Marx' 'menschliche Emanzipation' (1974).
7. If, as many argue, a PD can represent Hobbes' state of nature, it is a state of order, albeit a 'natural order' as explicated here.
8. This property of a strategy profile can be understood as an enhancement of a property introduced by Davis Lewis to characterize the stability of a convention. A 'coordination equilibrium' according to Lewis (1969, 14) is a strategy profile such that no deviation from the profile is advantageous to any of the players involved, if all others stick to the given. Thus, in a coordination equilibrium no actor has an interest in any other actor deviating from the common scheme, so long as no other does. The distinguishing feature of the simple coordination problem in Fig. 2 is stronger: every isolated deviation from the common scheme by one actor is strictly disadvantageous to every actor involved. Thus every actor has an interest that no singular actor deviates. A profile with this property could be called 'strict coordination equilibrium'.
9. Michael Tomasello argues that this is what decisively discriminates human beings from non-human primates. See, for example, Tomasello (2009, 32 f. 62 ff).
10. In what follows I will refer to the Selby-Bigge edition of Hume's Treatise (1978) by a capital T usually followed by the relevant page numbers. Similarly E will refer to Hume's Enquiries (1975).
11. See FN 8.
12. A prominent example put forward by Hardin is the electoral system in the USA. See Hardin (2014, 83 f).
13. See, for example, Hardin (2007, 92, 97, 98); or Hardin (2014, 87, especially FN 11).

14. It may, for example, be immoral if the rule represents a 'strict' coordination equilibrium and no other can be expected to act according to the rule. This, of course contradicts Kant's claim that the categorical imperative (as given in the first definition, see Kant 2002, 37) has unconditional obligating force.
15. I take it that these two claims about the explanatory and justificatory force of consent are the essence of the anti-contractarian argument of the dual coordination theory. See also the related short discussion in section 'Rules as Reasons for Action'.
16. For a more thorough discussion of moral approbation in the context of artificial virtues, see Lahno (1995, chap. 8).
17. According to Hume's analysis common sense morality tends to be deontological rather than consequentialist. Kant may well have been inspired by this piece of moral psychology. But, of course, no normative or meta-ethical conclusion can be drawn from this purely descriptive account.
18. The connection of this psychological claim to Leon Festinger's *Theory of Cognitive Dissonance* (1957) is obvious.
19. See Salter (2015) for a more comprehensive related argument.

References

Berlin, Isaiah. 1969. *Four Essays on Liberty*. Oxford: Oxford University Press.

Binmore, Ken. 2005. *Natural Justice*. Oxford/New York: Oxford University Press.

Constant, Benjamin. 1819. *The Liberty of Ancients Compared with That of Moderns*. Online Library of Liberty, Liberty Fund. http://oll.libertyfund.org/titles/2251. Retrieved 4 Sept 2015.

de Jasay, Anthony. 2010. Ordered Anarchy and Contractarianism. *Philosophy* 85: 399–403.

Ferguson, Adam. 1995 [1767]. An Essay on Civil Society, ed. Fania Oz-Salzberger. Cambridge: Cambridge University Press.

Festinger, Leon. 1957. *Theory of Cognitive Dissonance*. Stanford: Stanford University Press.

Hardin, Russell. 1988. *Morality Within the Limits of Reason*. Chicago/London: The University of Chicago Press.

———. 1989. Political Obligation. In *The Good Polity: Normative Analysis of the State*, ed. Alan Hamlin and Philip Pettit, 103–119. Oxford: Basil Blackwell.

———. 1995. *One for All. The Logic of Group Conflict*. Princeton: Princeton University Press.

———. 1999. *Liberalism. Constitutionalism, and Democracy*. New York/Oxford: Oxford University Press.

———. 2004. *Indeterminacy and Society*. Princeton/Oxford: Princeton University Press.

———. 2007. *David Hume: Moral and Political Theorist*. New York/Oxford: Oxford University Press.

———. 2014. Social Yes; Contract No, In Lahno & Brennan 2013/2014, 79–93. *RMM* 5. http://www.rmm-journal.de/downloads/Article_Hardin.pdf. Retrieved 16 May 15.

Hume, David. 1975. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. L.A. Selby-Bigge, 3rd ed., with text revised and notes by P.H. Nidditch. Oxford: Clarendon Press.

———. 1978. *A Treatise of Human Nature*, ed. L.A. Selby-Bigge, 2nd ed., with text revised and notes by P.H. Nidditch. Oxford: Clarendon Press.

Kant, Immanuel. 2002. *Groundwork for the Metaphysics of Morals*. Trans. and ed. Alan W. Wood. New Haven/London: Yale University Press.

Lahno, Bernd. 1995. *Versprechen. Überlegungen zu einer künstlichen Tugend*. München: OIdenbourg.

Lahno, Bernd, and Geoffrey Brennan eds. 2013/14. *Can the Social Contract Be Signed by an Invisible Hand?* Special Topic of RMM, Vol 4/5 2013/14. http://www.rmm-journal.de/htdocs/st03.html. Retrieved 10 Aug 15.

Lewis, David. 1969. *Convention*. Cambridge: Harvard University Press.

Marx, Karl. 1974. Zur Judenfrage. In *Marx Engels Werke (MEW) 1*, 347–377. Berlin: Dietz. English in: *Selected Essays* (trans: Stenning, H.J.). London/New York 1926: Leonard Parsons, 40–97.

Rousseau, Jean-Jacques. 1987. *The Basic Political Writings*. Trans. and eds. A.C. Donald, P. Gay. Indianapolis/Cambridge: Hackett.

Salter, John. 2015. Hume Without Spontaneous Order. In Lahno & Brennan 2013/2014, 26–38. *RMM* 6. http://www.rmm-journal.de/downloads/Article_Salter.pdf. Retrieved 16 May 15.

Sugden, Robert. 2009. Can an Humean be a Contractarian. In *Perspectives in Moral Science*, eds. Michael Baurmann, Bernd Lahno, 11–23. Frankfurt: Frankfurt School Verlag. Also *RMM* 0. http://www.rmm-journal.de/downloads/002_sugden.pdf. Retrieved 10 Aug 15.

Tomasello, Michael. 2009. *Why We Cooperate*. Cambridge MA/London: MIT Press.

Vanberg, Viktor, and James M. Buchanan. 1988. Rational Choice and Moral Order. *Analyse und Kritik* 10: 138–160.

Russell Hardin's Hobbes

Paul-Aarons Ngomo

Preliminary Remarks

Perusing Russell Hardin's corpus one cannot but be struck by the growing presence of Hobbes in his impressive body work in the years immediately following the publication of *Morality Within the Limits of Reason* (Hardin 1998). There, Hobbes is entirely absent from his "reconstruction of utilitarianism" as it relates to "the problem of choosing in social life" (Hardin 1998: ix). Likewise, apart from two passing references, his book on *Collective action* does not extensively discuss Hobbes's thought. The first reference appears in a brief survey of historical instances of the logic of collective action "stated by numerous political philosophers and political economists from the time of Hobbes to that of Sidgwick and Pareto, with especially elegant examples of the problem invented by Hume and J.S. Mill"; the second is a brief evocation of Hobbes's justification of coercion in a short segment on the efficacy of sanctions to make one-shot contracts workable.[1]

Contrastingly, a turning point occurs with "Constitutional Political Economy: Agreement on Rules", "Why a constitution?", and "Contractarianism: Wistful Thinking", three seminal papers that typify what we might characterize as Russell Hardin's Hobbesian turn (Hardin 1988, 1989, 1990). In these pivotal papers that lay out his arguments

P.-A. Ngomo (✉)
African School of Economics, Abomey-Calavi, Benin

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_5

against contractarianism and the core of his views on constitutions as coordinating devices, references to Hobbes are virtually ubiquitous. From there onward, his use of Hobbes's insights will only grow in importance to become a permanent fixture in his work. Even in *One for All: The Logic of Group Conflict* (1995), his penetrating work on violent conflict bears the imprint of Hobbesian intuitions. They are particularly noticeable in his analysis of centrifugal mobilization spurred by group coordination in the pursuit of group-level benefits.

In such contexts, group coordination often floods institutional barriers to violence by escalating conflict. The outcome is often what Hardin describes as a "structured variant of the state of nature" (Hardin 1995: 9). Once violence is unleashed, "preemption becomes an unavoidable urge. One need not hate members of another group, but one might still fear their potential hatred or even merely their threat. Hobbes's vision of the need of all to preempt lest they be the victims of the few who are murderous still fits even in the relatively organized state of ethnic conflict, except that it applies at the group level" (Hardin 1995: 144). Hobbes's presence in Hardin's work is perhaps nowhere as prominent as in *Liberalism, Constitutionalism, and Democracy*, arguably his major work (Hardin 1999a: 379). While canvassing what he takes to be his central argument, he argues that "Hobbes supposed it clearly serves our mutual advantage to have the most draconian government rather than to live in anarchy or civil war" (Hardin 1999a: 4).[2] Subsequently, we are told that "with better parlour-room manners in their discourse and a generous addition in their arguments, it is a thesis shared by the *Federalist Papers*, Tocqueville, and such contemporary democratic theorists as Robert Dahl" (Hardin 1999a: 4). Conjoining such diverse thinkers allows Hardin to locate his work within a philosophical lineage that stretches back to Hobbes, albeit with a few caveats, especially "the generous addition" stipulating that "if a society can coordinate on basic political and economic order, then it can risk politics at the margin over lesser issues. . .Where there is broad consensus on order, we do not need Hobbes's autocrat to rule us". Once the caveats are stated, pride of place is fully granted to Hobbes as a source of enduring insights useful in refining a theory of order to illuminate the structure of strategic interactions. As Hardin writes, "with that addition and one qualification, or emphasis, *Hobbes's thesis is also the thesis of this book*. The qualification, addressed especially . . .is that in some societies there is little hope of coordination on mutual advantage—conflict is too divisive and beyond compromise" (Hardin 1999a: 5, italics added). We would certainly miss the

impetus behind Russell Hardin's mid-career frequent evocations of Hobbes if we construe them simply as an erudite foray into the history of political thought in search of compelling illustrations to buttress arguments.

He turns to Hobbes while reconstructing the genealogy of theories of order in different historical settings to highlight various attempts to clarify the nature of the problem of order. In so doing, he identifies a thematic continuity across traditions and issues by locating past thinkers in a broader lineage that displays increasing levels of conceptual sophistication. This is especially apparent in Hardin's characterization of Hobbes as a "nascent coordination" theorist of order whose solution to the central explanatory problem of empowering a sovereign is subsequently improved by Hume's better grasp of the nature of iterated interactions and their institutional implications (Hardin 1999a: 13). We now begin to fully understand what is at stake in Hardin's Hobbesian turn. The bifurcation happens when central Hardinian themes have already been laid out in *Collective action* and *Morality Within the Limits of Reason*. Indeed, familiar topics such as Hardin's view of "coordination for mutual advantage" in an iterated coordination game—to achieve a "convention" feature prominently in the first of these books. In a similar vein, the latter fleshes out an account of institutional utilitarianism discernible behind his view that the "early origins of the general utilitarian justification of government" may be found "in the theory of Hobbes" (Hardin 2003: 12).

The supposition that Hobbes justifies government in utilitarian terms might seem rightly anachronistic, not least because such terminology postdates his era. This rather liberal use of adjectives conveys the general tenor of Hardin's reading of Hobbes. Eschewing exegesis, he is primarily interested in mining strands of ideas arising in scattered places in his justification of government to make them more intelligible to contemporary readers. To put it differently, he revisits Hobbes to retrieve modal categories that fit the underlying structure of a variety of strategic interactions. His goal is "to join the enterprise of re-reading Hobbes as a proto-game theorist" (Hardin 1991: 19). The exercise yields a recasting of Hobbes as a pioneering figure in the rise of the strategic analysis of dyadic or group-level interactions. In one such instance, he is praised as "the first major coordination theorist" (Hardin 1999a: 11). Elsewhere, he is credited with laying "the foundations for rational choice theory and for a major branch of political philosophy thereafter" (Hardin 2007: 208).

But Hardin's reading of Hobbes is not merely laudatory. His intent is perhaps best expressed in the words of Gregory Kavka, another prominent

Hobbesian revisionist, who is adamant in his insistence that "if Hobbes's philosophy is to be taken seriously today, it must be modified in certain respects. Some of his arguments must be modified or discarded" (Kavka 1986: xii). Endorsing just such a program, Russell Hardin revisits Hobbes to assess his central insights in light of subsequent developments in explanatory discussions on the structure of workable constitutional orders. In the end, he offers an interpretation that appropriates and redeploys Hobbesian insights often in ways that may be deemed unorthodox. Anticipating that his interpretive audacity might elicit skeptical gazes, he readily concedes, referring to the characterization of Hobbes as coordination theorist, that "not all scholars would agree with this assessment" (Hardin 1999a: 11 fn 23). This candid concession is indicative of the purpose of his rereading of Hobbes.

Approaching Hobbes as a theorist of order, he reads him as a thinker who was only partly successful because his tools and his general inclination left him unable to fully grasp the strategic characteristics of iterated interactions. As a result, he is unable to understand how conventions may arise from such interactions to stabilize expectations to render draconian enforcement superfluous. Though the Hobbes that emerges from Hardin's probing and reconstructive grip might seem slightly unusual, he is confident that "Even someone, at least a political philosopher, who disagrees with my account of him should nevertheless be interested in the theory I attribute to him, because it is a wonderfully spare baseline theory" (Hardin 1999a: 2 fn 6). The theory imputed to Hobbes purports to explain how order is maintained once it arises from dyadic interactions. Hardin finds it compelling enough to endorse it unreservedly as a self-standing theory should anyone cast doubt on its Hobbesian credentials. As he contends, "It is not important . . .whether my account of Hobbes is correct- those who think I have Hobbes wrong can read my account of his theory as my theory of social order" (Hardin 2001a: 65).

Whether Hardin should retain the theory imputed to Hobbes as his own is not my central concern here. Instead, my goal is partly expository and partly evaluative. I wish to sketch a reception of his critique, appropriation, and creative redeployment of Hobbesian insights. The purpose of such a reception is to highlight the varied aspects of his rereading of Hobbes as he reconstructs the structure of Hobbes's arguments to determine their epistemological status in a broader lineage of social-scientific thinking on political order. The result, I suggest, is a *critique* of Hobbes, that is, an examination of the possibilities and limits of the conceptual framework that grounds his theory of government. Briefly described, Russell Hardin's

interpretation of Hobbes revolves around three related arguments. First, he is presented as the progenitor of a minimalist value theory built on a "holistic normative principle" that justifies mutually advantageous institutions (Hardin 2003: 121). In contrast to standard deontological theories of political justification, he is said to subscribe to a welfarist vision of order derived uniquely from self-interest with no prior normative commitment. Second, Hobbes is praised for "seeing the two-level nature of our general problem of political justification". As Hardin would have us believe, "This is an aspect of Hobbes's account that has received inadequate recognition by subsequent thinkers" (Hardin 1999b: 105). Accordingly, an adequate recognition requires identifying the methodological insight that makes Hobbes appealing even when we disagree with his normative conclusions, as Hardin does. He contends that Hobbes's major insight lies in "establishing a two-stage theory of government and of justification. At the first stage, we create and justify government. At the second stage, government creates policies" (Hardin 1999b: 106). Finally, Hobbes's contractarian justification of institutions is rejected as a "lousy theory" (Hardin 2007: 87 fn 14) that misframes the structure of the problem of maintaining government. In Russell Hardin's uncharacteristically uncharitable assessment, "His story is fundamentally silly and of no real interest" (Hardin 2007: 81). The contractarian story is allegedly unworkable because individuals cannot literally transfer their natural faculties or their instrumental power to a sovereign. Overall, Hardin boldly claims, Hobbes is hamstrung by his contractarian theory of the creation of government and somehow errs when he envisions individuals involved in the initial empowerment of government as parties facing a "one-time coordination problem" (Hardin 2007: 56).

Because he supposedly misdescribes the strategic structure of the problem of order, Hobbes prescribes draconian enforcement to solve the problem of the maintenance of government. His "gunman view of the sovereign" empowers a government endowed with unconstrained power to keep subjects in line (Hardin 1991: 159). This "institutionally enforced coordination" is deemed "empirically wrong" because it overlooks the stabilizing potential of repeated interactions for social cooperation (Hardin 2011: 45). In trying to empower government, Hardin argues, agents are faced with a repeated coordination problem, not one that could be resolved satisfactorily with an expedient designed for one-time interactions. While I am generally sympathetic to Hardin's enterprise, I wish to suggest that his revisionist interpretation is at times so liberal as to raise concerns

about its orthodoxy. Because Hardin often explains Hobbes's views by contrasting them with those of Hume, his humeanized Hobbes is often dehistoricized. His chronocentric bias leads him to see in Hobbes a failed theorist of order who could not grasp the structure of iterated coordinations. In doing so, he misses Hobbes's central insight about the efficacy of a prudent exercise of sovereignty to achieve the maintenance of order.

In what follows, I examine Russell Hardin's Hobbes in three steps. The next two sections ("Hobbesian Value Theory" and "The Coordination Theory of Order: Hobbes in Russell Hardin's Humean Mirror") are primarily expository. They focus on Hardin's reception and modification of Hobbes's insights on the maintenance of order. In the section "The Sovereign at the Helm", I argue that his convention-based account of the maintenance of order runs against Hobbes's view that collective prosperity depends on a successful "exercise of entire Soveraignty" (Hobbes 2014: 574), and with one that acknowledges the centrality of the sovereign in the maintenance of order. On this contrasting account, though the sovereign embodies an enforced coordination, continuous order depends on how she discharges the duties attached to her office to procure the "safety of the people" and "also contentments of life", rather than "a bare preservation" (Hobbes 2014: 520). In short, governance, not endogenously emerging conventions, incentivizes cooperative behavior and acquiescence to extant order.

Hobbesian Value Theory

The claim that Hobbes's justification of government "starts from a particular value theory" hardly features in standard accounts of his moral theory (Hardin 2001b: 60). They are typically preoccupied with the normative status of his account of the laws of nature. A great part of the discussion revolves around whether they are best interpreted as deontological principles, divine commands, or simply as prudential precepts.[3] As a result, it has very little to say about the value theory that grounds the strategic move to endorse order. In his interpretation of Hobbes's justification of government, Hardin provides an account of cooperation that rejects visions of moral obligation derived from an overarching theory of the good or the right. Unlike "a fairly standard moralized account of Hobbes's intent" that articulates "a theory of moral obligation to the sovereign that follows merely from contracting" (Hardin 1999a: 19–20), he reads Hobbes as expounding

a utilitarian rather than a deontological theory of government that grounds decisions on some exogenous moral principle.

We choose life under government not because of its intrinsic goodness, but primarily insofar as it is the most efficient means to achieve desired ends. "Government has no value in its own right, it is merely the means to the end of human welfare", Hardin argues (Hardin 1999a: 47). While he generally endorses psychological explanations of motivation that subordinate reasons for action to self-interest, he does not follow them in presenting a prudential account of action that describes Hobbes's enterprise as one whose "primary aim is to demonstrate what men ought, and what they ought not, to do", as David Gauthier contends (Gauthier 1969: 27). Rather than articulating a prescriptive theory of human motivation, Hardin presents a *Hobbesian value theory* that purports to illuminate the motivational basis of mutually agreeable commitments. This claim should not be taken as implying that Hardin's aim is to articulate a self-standing value theory that takes its bearings from Hobbes to explain, "in a general way, why morality exists, why it has the content it does", as Kavka does in his "Hobbesian analysis of morality" (Kavka 1986: xiv).

Though he often uses a terminology that wasn't available in Hobbes's time, his discussion focuses entirely on clarifying the normative underpinnings of Hobbes's axiology. Like Hobbes, he rejects the inherentist view that value inheres in things. The central claim of his account is that "Hobbes's value theory was individualist and ordinalist" (Hardin 2003: 43). While describing Hobbes's premises as individualist fits standard interpretations of his account of human motivation, it is less common to cast him as an ordinalist value theorist. The claim appears all the more anachronistic since such a term was not part of his philosophical lexicon. But it conveys a central intuition of Hobbes's holistic justification of a particular state of affairs as ordinally better than the contrary choice. The valuation is ordinal insofar as the choice is between two states of affairs with contrasting implications. We may say that "Hobbesian individualist ordinalism" grounds valuations that elicit choices that meet individual expectations holistically since they are typically "focused on the problem of collectively providing for individual welfare" (Hardin 2007: 173). Furthermore, Hardin argues, "the central move of such theory is typically to create an institutional structure that will guarantee the welfare of individuals who act sensibly, which is commonly to say, who act according to the simple canons of rational choice" (Hardin 2007: 173). This claim needs unpacking. There are two related points at play here. First, valuations are a reflection of

individual expectations in the following sense: my choice of a state of affairs as the most desirable is driven by the urge to obtain conditions conducive to my welfare. Second, my choice is only cashed out collectively since my preferred state of affairs coincides with the expectation of other agents striving to achieve their welfare. The baseline that allows each of us to secure our welfare is mutually advantageous because it works aggregatively. In this sense, "we may say that mutual advantage is the collective implication of self-interest because to say that an outcome of our choices is mutually advantageous is to say that it serves the interest of each and every one of us. One could say that, in this view, collective value is emergent; it is merely what individuals want" (Hardin 2003: 14).

On the preceding view, the emergent collective in value is order. We are all strictly comparatively better off under orderly government because the rise of order is a holistic resolution that saves us all from the nefarious prospects of a "solitary, poore, nasty, brutish, and short" (Hobbes 2014: 192) life that awaits in a world in which we are left to our own devices to fend for ourselves, each of us assured of a dreadful untimely demise. Hence, an ordinalist valuation grounds Hobbes's justification of orderly government. For Hardin, "A striking feature of Hobbes's view is that it is a relative assessment of whole states of affairs: Life under one form of government versus life under another or under no government at all" (Hardin 2003: 43–44). Individual values are realized collectively since it is in our mutual interest to endorse an institutional structure that stabilizes expectations. As such, crediting Hobbes with the "astonishing success of founding government in an account from self-interest, or rather from its collective implication in mutual advantage" amounts to stressing that collective welfare is essentially self-interest writ large.

Accordingly, a theory of the good or the right that derives its normative force from an abstract moral impetus that requires us to act justly does not yield workable valuations because it fails to grasp that collective resolutions are reliably efficient when they speak to individual valuations and expectations about welfare, that is, when they speak to our self-interest and motivate us on such grounds. This is Hardin's ordinalist interpretation of Hobbes's value theory. Recall that he is primarily interested in identifying the normative principle that aligns individual valuations, that is, self-interest, with collective concerns, so that resolving the latter only requires a "motivational theory" predicated on "disaggregated individual values" (Hardin 2003: 13). The casualty of this rereading is the deontological interpretation of laws of nature as transcending norms that motivate action for purely

moral reasons. In Hardin's account of Hobbesian value theory, valuations are self-centered and primarily endogenous since only circumstances of choice and their constraints determine how individuals act, regardless of other religious or moral commitments they may have. Facing the ordinalist choice between order and deadly anarchy, individuals would prefer the former because it comports with their fundamental values that motivate them, namely "survival and welfare" (Hardin 2003: 13).

To sharpen the contrast between his minimalist account of Hobbesian value theory and deontological views of the laws of nature, Hardin boldly reinterprets them as "*sociological laws about what would work to our interest*, not because they are in some sense moral" (Hardin 1999a: 2). Hence, we might say that his view of Hobbes's value theory turns self-interest into a normative principle that provides a baseline to motivate collective resolutions of individual problems. What is more, endogenizing laws of nature by framing them as workable precepts rather than as exogenous commands, as is often the case in deontological readings, offers an insightful interpretation of the connection between Hobbes's moral and psychological claims and his political theory. The primacy of order implies that moral concerns are only derivatively important, that is, when they fit the structure or order, as is the case when normative orders match self-interest to elicit cooperation without "an ad hoc claim of normative commitments" (Hardin 2003: 54). In Hardin's reading of Hobbes's value theory, the values of survival and welfare ground stable orders. It is the only normative foundation they require to achieve stability. Hence, the claim that "Hobbes's theory of government required no normative principle of obligation" reaffirms the primacy of self-interest as a minimalist principle of valuation. The positive and the normative are inescapably intertwined because they are driven by "related motors", that is, "individual incentives for individual benefits" (Hardin 2001b: 61).

Let us pull together the preceding claims to summarize Hardin's account of Hobbes's value theory. Because self-interest drives our valuations, our commitments are unlikely to be shaped by exogenous normative stances. With the exceptions of a fringe of religious fanatics or glory-seekers in our midst, survival and welfare typically justify our commitments, especially acquiescence to orderly government in the face of the possibility of chaos. This ordinalist claim grounds Hobbes's view of political justification. Despite claims to the contrary that interpret him as arguing "that because we have agreed to government we are morally obligated to stick by our agreement", Hardin characterizes his "actual justification" as "more nearly

utilitarian because it is grounded in mutual advantage". Insofar as "we are all better off to have a state and, once we have one, to avoid dissension and revolution", the value theory that underlies the justification of government is fundamentally welfarist, not deontological (Hardin 2001b: 69). Valuations predicated on the values of survival and welfare turn mutual advantage into "relatively compelling holistic normative principle" because they elicit patterns of coordination that might generate positive implications for the whole society. Hardin calls "Hobbesian efficiency" the grounding of "value theoretic accounts in individuals" to justify collective choice. Its appeals lies in providing a holistic justification for order or the structure of a legal system (Hardin 1993a: 463). This is the foundation of Hardin's claim that "Hobbes's theory of political sovereignty has its minimalist moral grounding in mutual advantage" (Hardin 1993b: 362). Accordingly, Hobbes is at once "welfarist and resourcist" since prospects of "greater welfare" in a stable order are enticing enough to elicit commitment to stave off chaos. Perhaps the most striking feature of Hobbesian value theory as Hardin reconstructs it is that it is essentially political and not a separate moral theory. To put it slightly differently, Hobbes derives the structure of institutions "at the large scale of the whole society" (Hardin 2007: 107) from individual valuations instead of predicating them on a moralizing view of political action. His value theory is not strictly a moral theory that articulates a moral vision of the political world. It is a unified theory of political justification that connects personal valuations with the structure of institutions that holistically secure survival and welfare. This is Hardin's central claim in his interpretation of Hobbes as a coordination theorist. It is examined in greater details in the next section.

The Coordination Theory of Order: Hobbes in Russell Hardin's Humean Mirror

The preceding discussion on Hobbesian value theory has set up the revisionist rereading of his account of political order. As we have seen, a normative principle derived from our valuations grounds our choice of institutions. Order arises and remains stable because it meets our expectations about survival and welfare. While Hobbes is clear about the causal role of these foundational values in stabilizing political order, he is far less explicit about the structure of his theory of political justification, Hardin contends. In Chapter 17 of his *Leviathan*, his arguments are couched in a

contractarian idiom that obscures the motivational basis of acquiescence. There, he argues that political unity is "made by covenant of every man with every man" to empower government (Hobbes 2014: 260). Hardin discards his contractarian metaphor to recast him as a theorist of mutual advantage. He is said to have articulated "the clearest, most urgent claim for the mutual advantage of orderly government" to justify an institutional structure that secures cooperation and allows people to seek economic prosperity on their own terms (Hardin 1999a: 4). Whereas the contractarian just-so story about the creation of government figuratively involves parties in a starting position agreeing "to conferre all their power and strengths upon one Man, or upon one Assembly of men, that may reduce all their Wills, by plurality of voices, unto one Will", mutually advantageous acquiescence is striking in that it does not imply an argument between parties figuratively or literally (Hobbes 2014: 260). All that is required is merely a coordination that secures an outcome that suits the most significant segments of the relevant population and compels potential dissenters to go along with the arrangement because failure to comply might potentially leave them worse off.

Russell Hardin starts his assessment of Hobbes's account of political order by praising him as a remarkable proto-game theorist and a proto-utilitarian who "justifies a government according to the benefits it offers relative to alternative governments" (Hardin 1991: 175). He locates his main insight in his theory of government that distinguishes between *ex ante* decisions and those taken, *in media res*, that is, under mutually advantageous institutions, once they are established. In this account ascribed to Hobbes, he is said to have articulated a determinate resolution of the problem of choosing and justifying institutions. In Hardin's assessment, "He did so by establishing a two-stage theory of government and of justification. At the first stage, we create and justify government. At the second stage, government creates policies" (Hardin 1999a: 106). It is suggested that Hobbes's central methodological insight also finds an echo in various resolutions of the problem of institutional justification of particular choices in many other settings. A conceptual genealogy of mutual advantage puts Hobbes in impressive company. Presumably, the normative structure of the two-stage theory also grounds liberalism, democracy, and constitutionalism (Hardin 1999a: Chap. 1). Likewise, Rawls whose focus is "on the general structure of political-legal order" is squarely placed in a lineage of mutual advantage theories that counts Hobbes among its foundational figures (Hardin 2003: 103). Hardin goes so far as to argue that

"the normative foundations of the formulations of Hobbes, Pareto, and Coase are essentially the same" (Hardin 1996, 1991).

The reconsideration of the structure of Hobbes's arguments begins with a startlingly heterodox claim. In contrast to the received view that sees him as a consent theorist of order because he spoke of a covenant to establish government, Hardin rejects the very idea of a consensus between consenting parties agreeing to divest themselves of their natural and instrumental powers to set up orderly government. He suggests that Hobbes lacks clarity because he is often "ambivalent about what problem he wishes to resolve. His discussion is more or less equally about the creation and the maintenance of sovereign government" (Hardin 1991: 157). Hobbes, it is suggested, blurs the distinction between two separate issues: creation and maintenance of government. His ambivalence is apparent in the contrast between his contractarian account of the creation of order and his generally welfarist views of its maintenance. His theory of the creation of a sovereign by institution sits uneasily with his account of sovereignty by acquisition. Indeed, his convoluted justification of the power of the sovereign by acquisition does little to dispel the suspicion that sovereignty obtained through conquest is hardly defensible on a contractarian basis. In such realms, compliance is obtained by coercion -not by consent- since the defeated and subjugated populace is, de facto, powerless before its powerful conqueror. Moreover, the contractarian story is arguably a "lousy theory because it runs against the strategic problem of transferring power from all individual citizens to the sovereign and it is historically irrelevant" (Hardin 2007: 87). Likewise, Hardin says, "the theory of maintenance requires only Hume's convention for order and Hobbes does not give us an account of how ongoing order works" (Hardin 2007: 87). The rejection of the contractarian theory of order goes hand in hand with a restatement of the problem of maintaining government without Hobbes's overbearing sovereign. Before unpacking the convention-based theory of the maintenance of government, I briefly present below the main argument leveled against the contractarian account of the empowerment of government.

To begin, recall that all that is required to maintain government is to align self-interest with extant order. Indeed, the centripetal force of mutually advantageous institutions stems from their ordinal superiority over competing alternatives. A comparative assessment of whole states of affairs (order vs. chaos) leads us to endorse the only alternative that fits our expectations about security and prosperity because it causally generalizes self-interest. To commit to extant order, we merely need to acquiesce to

government by submitting to its commands. In the contractarian variant of the initial creation of government, order arises following a transfer of power from consenting parties to a sovereign endowed with the prerogative to secure compliance with the power of the sword bestowed upon her. In Hardin's view, any such transfer is de facto impossible. In making this claim, Hardin says, "Hobbes falters". He writes, "We can consent all we want to but, as a matter of actual fact we cannot simply hand our power over to anyone if that power is constituted primarily of our human capacities. . .. I consent to the movement of the mountain before us out of our path, but it will not happen therefore. And our new sovereign cannot enter office with any power worth having for the awesome tasks ahead" (Hardin 2007: 215). The contractarian story is unworkable because individuals cannot literally transfer their natural faculties or their instrumental power to a sovereign. Hence, the very idea of a transfer of power is merely a derivation by *fiat* because it cannot produce the expected aggregation of individual strengths that supposedly establish an all-powerful sovereign.

Having established that "Hobbes lacks a credible account of the initial empowerment of the government" because the very idea of a transfer of power that sustains the contractarian view is unsalvageable, Hardin reinterprets him as a coordination theorist to spell out the structure of his account of political order in a much clearer light than what we typically see in standard contractarian justifications of government. On this view, Hobbes is best read as a coordination theorist because the initial issue of empowering government is about providing a holistic resolution to collective problems, not to elicit contractual commitments. While a contract addresses a collective action problem that arises when our resolutions may be undone by free-riders, an initiation empowerment of government faces no such issues. As Hardin explains, the establishment of government is not akin to tackling "an n-prisoner's dilemma that must be resolved by agreement and then enforcement of the agreement" (Hardin 2007: 110). In practice, "If we once do establish a sovereign, there is no prospect of free-riding on that choice. I might prefer to be able to avoid the sovereign's glare when I wish to steal from you in the political society on which we have coordinated. But I cannot free-ride on the initial coordination itself" (Hardin 2007: 110). Hardin's creative insight here is to reinterpret Hobbes by reframing his central problem as one of coordinating to maintain order, not one of contractually creating from the state of nature.

The recasting of Hobbes as a theorist of the maintenance of order is consistent with "his overriding actual concern at the time of writing, even

his likely motivation for writing". That concern was "the maintenance of sovereign government in the face of revolutionary fervor and turmoil" (Hardin 1991: 157). If Hobbes is not the contract theorist he is often made out be, including by none other than himself, it is because he clearly grasped the impossibility of a universal agreement to establish order, as it happens when a fraction of dissenters such as religions extremists or thrill seekers are not interested in securing their personal survival and prosperity. Hobbes says as much when he claims that "because the major part hath by consenting voices declared a Soveraigne; he that dissented must now consent with the rest; that is, be contented to avow all the actions he shall do, or else justly be destroyed by the rest" (Hobbes 2014: 268). This observation fits the strategic structure of resolutions that make most people better off in society. Hobbes's tacit view here is that extremists and those willing to disrupt order in pursuit of gains incompatible with the security and the welfare of significant groups should be overrun to preserve peace for those whose self-interest aligns with extant government. In this sense, he is best described as a coordination theorist.

Yet even as he recognizes Hobbes's credentials as a coordination theorist, Hardin is quick to emphasize that his limited grasp of the structure of the problem of maintaining order betrays analytical shortcomings he could not correct. Beholden to his contractarian idiom, he could not cut the Gordian knot to overcoming the crippling ambivalence between creation and maintenance of government. Hardin deems Hobbes's grasp of the problem of maintaining government insufficient because he saw it mainly as one of enforcement through "draconian force". Resolutions obtain though institutionally enforced coordination by getting everyone to select order over chaos. While his "one-time coordination" is resolved by empowering a sovereign, the resulting order requires political absolutism to remain stable, a state of affairs that makes changing government perilous. Hence, Hobbes's resolution is inherently conservative and biased toward the status quo since attempting regime change might wreak havoc on the prevailing coordination that preserves ongoing orderly interactions.

In a Hobbesian polity, subjects would be stuck with a powerful sovereign with no prospects for political liberalization because Hobbes sees the problem of maintenance as a one-time coordination. As a result, he fails to grasp how repeated interactions "could lead to very stable, compelling incentives for continuing coordination that is spontaneous and that is not deliberately organized through an explicit agreement or overseen by any manager to keep us in line" (Hardin 2007: 214). Herein lies the crux of Hardin's

reinterpretation of Hobbes as a coordination theorist. To redeem Hobbes, he turns to Hume to mine his insights on the institutional implications of iterated interactions to provide an account of large scale coordination for order without draconian enforcement. Reading Hobbes through humean lenses magnifies his analytical shortcomings and reveals the extent of his incomplete grasp of the challenging problem of maintaining government after the initial coordination to secure peace.

In humeanizing Hobbes, Hardin was not merely trying to dismiss his limited insights on strategic interactions. His move might be construed as a reconstruction that sheds light on the gradual improvement of our understanding of how order is maintained. Thus, from Hobbes's "one-time coordination" through Hume's more sophisticated account of iterated interactions, we can readily see a development that mirrors various steps of a broader conceptual transformation. It begins with Hobbes, journeys through Hume and reaches its peak with game-theoretic theorizing on strategic interactions. Using Hume's better grasp of the institutional implications of repeated interactions as a critical standpoint allows Hardin to show how ongoing order works, something Hobbes could not clarify perhaps owing to the limitations of his tools. Because he could not see the stabilizing potential of iterated interactions, he is oblivious to the normative force of conventions. Consequently, Hardin says, he "overestimates the need for an especially powerful state to regulate behavior" (Hardin 2007: 214).

In contrast, Hume sees the strategic possibility of iterated interactions and clearly understands that it is in everyone's best interest to cooperate to maintain ongoing exchange and preserve future benefits. Hence, the benefit reaped in reading Hobbes through Hume's lenses, as Hardin does, is that we gain far better insights into how to "enable government to enforce its will, both at the outset and thereafter" without the persistent threat of incurring the wrath of the sovereign (Hardin 2007: 214). The central point here is that conventions rather than draconian rule stabilize expectations and preserve extant order. Moreover, subsequent political liberalization happens because "the incentives that back conventions can partially control even political office holders, who can be constrained in ways that Hobbes did not grasp", Hardin claims (Hardin 2007: 224). Strikingly, the sovereign as a governor who deploys the tools of statecraft to secure compliance is entirely absent from the convention theory of order, as if only ongoing coordinations determine the quality of governance and its outcomes. I argue in the next section that Hardin's reliance on what he sees as

Hume's "solution to Hobbes's central problem" leads him to overlook the role of the function of governance is maintaining political order (Hardin 2007: 224).

The Sovereign at the Helm

Strikingly, the convention theory of political order downplays the impact of the exercise of sovereignty in maintaining order, as though spontaneously emerging conventions rather than the performance of the sovereign matters in explaining political stability. While Hardin's rereading of Hobbes through Hume's account of the rise of conventions illuminates the role of emerging norms in stabilizing order, it is not entirely clear that focusing solely on orderly iterated interactions between agents involved in ongoing exchanges fully explains how government is maintained. Likewise, the view that the maintenance of order in Hobbes's theory of government requires only draconian force should be qualified. Brian Barry wrote quite uncharitably "that it is a travesty to identify Hobbes with a position in which social order depends solely on 'draconian force'" (Brian 2010: 370). Travesty is perhaps not the right word to describe Hobbes's understanding of the exercise of power. But the observation raises a key point in suggesting that the preservation of government does not solely require the use of force, as Hardin seemingly contends. I wish to argue that Hobbes's sovereign may preserve government without constantly using draconian force.

While Hardin is certainly right that "Hobbes's argument for the necessity of draconian force seems empirically wrong for many societies", he overemphasizes the role of blunt force in securing order and tends to proceed as though Hobbes's sovereign preserves order only through stringent enforcement of laws (Hardin 2007: 81). The fact that Hobbes offers a more nuanced account of the maintenance of government is apparent in his claim that force alone is not enough to elicit compliance. He argues for a program of civic instruction to instill in subjects the legitimacy of the rights of sovereignty. "And the grounds of these Rights, "Hobbes writes," have the rather need to be diligently taught; because they cannot be maintained by any Civill Law, or terrour of legall punishment" (Hobbes 2014: 522). Notwithstanding his defense of absolutism, Hobbes understood that draconian force is insufficient to secure compliance. As Arash Abizadeh puts it, "Hobbes did not believe that any sovereign could ever wield enough coercive power to maintain order on that basis alone" (Abizadeh 2010: 116).

Hobbes's recognition of the limits of "draconian force" to maintain order is just the tip of a theory of governance that casts the sovereign as an astute ruler who often relies on subtle means to maintain order. Indeed, "the exercise of entire Soveraignty" requires a proactive art of government that limits the needs for a brute force to maintain order (Hobbes 2014: 574). Copious prescriptions in his *Leviathan* offer guidance as to how the sovereign should behave at the helm. For example, Hobbes insists that rulers should not "countenance anything obliquely which directly they forbid" because "The examples of princes, to those that see them, are, and ever have been, more potent to govern their actions than the laws themselves" (Hobbes 2014: 476). Similarly, Hobbes notes that "power is preserved by the same Vertues by which it is acquired; that is to say, by Wisdom, Humility, Clearnesse of doctrine, and sincerity of Conversation; and not by suppression of the Naturall sciences, and of the Morality of Naturall Reason; nor by obscure Language" (Hobbes 2014: 1076). The larger implication is obvious: the exercise of sovereignty depends far less on draconian force than Hobbes's defense of absolutism suggests. We need to look no further than his recommendations on the strategic use of rewards and punishments to grasp the extent to which the maintenance of government depends on the skillfulness of the sovereign. A policy of selective incentives allows the sovereign to mete out punishments to offenders or show leniency in proportion to the dangers posed to the commonwealth by those who run afoul of the law. To a remarkable extent, the prescriptions on punishment and rewards afford us an insightful observational standpoint to understand the universe of governance and the challenging tasks a ruler inevitably faces once at the helm. Revealingly, when Hobbes points out that "the maintenance of civil society" depends "on Justice. . .and Justice on the power other lesse rewards and punishments" (Hobbes 2014: 698), he brings into full view the limits of draconian force in maintaining the commonwealth.

Concluding Remarks

Audacious interpretations of canonical thinkers unavoidably often raise questions of their own. Hardin's quip that whoever does not think his "account of Hobbes is correct" should read it as his is conceivably a subtle but resolute way to assert his Hobbesian credentials. But probing his claim to Hobbesian orthodoxy is unwarranted. Philosophical appraisal is not merely a drab exercise and redundant exercise in doctrinal loyalty. It is

only worthwhile when it derives new insights from paths often trodden with little creative impetus to express in other words claims that hardly need another superfluous restatement. On this score, Hardin's interpretation of Hobbes is largely successful despite conclusions that may seem controversial, a fact he readily conceded. In the preface to his book on David Hume, he writes that "when we read any theorist, and perhaps especially when a philosopher reads another philosopher, we often tend to take a strong critical stance and to pick the theorist apart" (Hardin 2007: vii). Drawing on Hume, Hardin reads Hobbes to harness insights that provide a strong starting vantage point to subsequent theorizing on order. For example, Hobbes supplies many of the conceptual structures underlying Russell Hardin's mutual advantage theory of social order, including his two-stage theory account of the empowerment that entails a broad *ex ante* agreement in the first stage to facilitate coordination of a workable order and subsequent resolutions that take place in *media res*. Rather than picking here and there bits and pieces of confirmatory wisdom, Hardin reconstructs Hobbes's intuitions and integrates them in his toolkit before improving them by revisiting them through humean and game-theoretic lenses.

His hermeneutical efforts highlight the novelty of Hobbes's ideas and their conceptual limitations as well. In reinterpreting Hobbes to clarify the "strategic structures of the problems that we face in achieving social order", Hardin provides a robust exemplar of how interpretive conversations across philosophical eras might improve our analytical toolkits (Hardin: 23). Paraphrasing Gregory Kavka, we may argue that his elucidation of the "modal strategic categories" that ground Hobbes's analytical reasoning embodies an approach that shows how a classic text can be creatively appropriated to contribute to contemporary philosophical debate (Kavka 1986: xiii). Perhaps unsurprisingly, the Hobbes that surfaces from Hardin's searching assessment is both flawed and brilliant. The flawed Hobbes is just as edifying as the brilliant one. His flaws show the limits of a social contractarian idiom that leaves him unable to expound a compelling theory of the maintenance of government. Once stripped of his contractarian garbs, the brilliant Hobbes emerges as theorist of power, that is, as a social scientist. In Hardin's assessment, "the difference between Hobbes the contract theorist and Hobbes the power theorist is the difference between a political philosopher and a social scientist" (Hardin 2006: 299). If his Hobbes often seems like a pared down version of Hume with a weaker grasp of the strategic structures of the problem of maintaining order, the comparison is not meant to be unflattering. It stands out as a perceptive

genealogical reconstruction that shows how early generations of thinkers provide the seeds that allow others to chart new territories. Undoubtedly, we are indebted to Russell Hardin for his bold interpretation of Hobbes as a thinker whose two-stage theory of order is far closer to contemporary concerns than it might appear at first glance.

Notes

1. See Hardin (1982). For the lone reference to Hobbes, see p. 8.
2. According to Hardin, "the Hobbesian view seems to fit ethnic conflicts that have turned violent in Lebanon, Azerbaijan and Armenia, Rwanda and Burundi, Iraq, and many other societies, as it fits Yugoslavia" (1999a: 144).
3. This simplification glosses over nuances. For a fuller account, see Martinich (1992, Chap. 3). For a more recent interpretation of the laws of nature, see Lloyd (2009).

References

Abizadeh, Arash. 2010. The Representation of Hobbesian Sovereignty. Leviathan as Mythology. In *Hobbes Today Insights for the 21st Century*, ed. S.A. Lloyd. New York: Cambridge University Press.

Barry, Brian. 2010. David Hume as a Social Scientist. *Utilitas* 22 (4): 369–392.

Gauthier, David. 1969. *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes*. Oxford: Oxford University Press.

Hardin, Russell. 1982. *Collective Action*. Baltimore: The Johns Hopkins University Press.

———. 1988. Constitutional Political Economy: Agreement on Rules. *British Journal of Political Science* 18: 513–530.

———. 1989. Why a Constitution? In *The Federalist Papers and the New Institutionalism*, ed. Bernard Grofman and Donald Wittman, 100–120. New York: Agathon Press.

———. 1990. Contractarianism: Wistful Thinking. *Constitutional Political Economy* 1 (2): 35–52.

———. 1991. Hobbesian Political Order. *Political Theory* 19 (2): 156–180.

———. 1993a. Efficiency. In *A Companion to Contemporary Political Philosophy*, ed. Robert E. Goodin and Philip Pettit. Malden: Blackwell Publishers.

———. 1993b. Altruism and Mutual Advantage. *Social Service Review* 67 (3): 358–373.

———. 1995. *One for All. The Logic of Group Conflict*. Princeton: Princeton University Press.

———. 1996. Magic on the Frontier: The Norm of Efficiency. *University of Pennsylvania Law Review* 144: 1987–2020.
———. 1998. *Morality Within the Limits of Reason*. Chicago: The University of Chicago Press.
———. 1999a. *Liberalism, Constitutionalism and Democracy*. New York: Oxford University Press.
———. 1999b. Deliberation: Method Not Theory. In *Deliberative Politics. Essays on Democracy and Disagreement*, ed. Stephen Macedo. New York: Cambridge University Press.
———. 2001a. Law and Social Order. *Philosophy Issues* 11 (1): 61–85.
———. 2001b. The Normative Core of Rational Choice Theory. In *The Economic World View: Studies in the Ontology of Economics*, ed. Uskali Mäki. New York: Cambridge University Press.
———. 2003. *Indeterminacy and Society*. Princeton: Princeton University Press.
———. 2006. Constitutionalism. In *The Oxford Handbook of Political Economy*, ed. Barry R. Weingast and Donald E. Wittman. New York: Oxford University Press.
———. 2007. *David Hume: Moral and Political Theorist*. Oxford: Oxford University Press.
———. 2011. Normative Methodology. In *The Oxford Handbook of Political Methodology*, ed. Robert Goodin. Oxford: Oxford University Press.
Hobbes, Thomas. 2014. In *Leviathan*, ed. Noel Malcolm. Oxford: Oxford University Press.
Kavka, Gregory. 1986. *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press.
Lloyd, S.A. 2009. *Morality in the Philosophy of Thomas Hobbes Cases in the Law of Nature*. New York: Cambridge University Press.
Martinich, A.P. 1992. *The Two Gods of Leviathan: Thomas Hobbes on Religion and Politics*. New York: Cambridge University Press.

Constitutions as Conventions: A History of Non-reception

Andrew Sabl

In 1983, Michael Harrington wrote a book called *The Politics at God's Funeral*. In that book, Harrington accepted the decline of faith but thought it left a spiritual void that he hoped a form of socialist politics could fill (Harrington 1983).[1] This paper treats, in a sense, the academic politics at convention theory's christening (or bris).

Over several books and related articles, culminating in his reconstruction of David Hume's political thought, Hardin expounded an account of constitutionalism that saw constitutional structures as profound and pervasive solutions to coordination problems.[2] David Hume first called the relevant kind of long-term and society-wide solutions to coordination problems *conventions*, and one may describe the whole theory, for short, as the convention theory of constitutions. Hardin's convention theory of constitutions implies several conclusions that fulfill the usual requirements for deeply influential social science by being striking, deep, profoundly

I would like to thank the participants at the Russell Hardin Festschrift conference, New York University, November 6–7, 2015. Special thanks are due to Jerry Gaus, Bernd Lahno, and Bernard Manin.

A. Sabl (✉)
University of Toronto, Toronto, ON, Canada

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_6

counterintuitive (at least when taken together)—and true. That is, constitutions simultaneously serve all those subject to them by enabling them to pursue more effectively their disparate purposes (more briefly, by serving their diverse interests). Constitutions promote citizens' interests unequally yet without the likelihood of provoking successful opposition. And they do so without requiring either explicit consent or normative consensus.

As far as I know, no one has directly refuted the heart of this theory. No one disputes that the members of large societies have an interest in following *the same* rules, modes, or norms of action—particularly with regard to the most fundamental political and economic structures—to some degree regardless of what the substance of those rules, modes, or norms will be. No one seriously believes that what I have elsewhere called "crude focal points," the kind characteristic of face-to-face societies (words within earshot, the grasping of a conch or a scepter in a gesture all can see, a reputation, known through personal acquaintance, for wisdom or decisiveness), can solve those problems when it comes to mass societies.[3] No one credibly denies, once it is pointed out, that the kind of legal and political conventions we call constitutions palpably do solve these kinds of problems much of the time, settling in particular which laws count as authoritative and how the holders of political power are to be selected.

But the truth, depth, and counterintuitiveness of Hardin's conclusions have not, alas, ensured their fame. My observation, which I'm afraid must remain an axiom for inability to prove a negative, is that while political theory—surely the natural home of Hardin's work—has never rebutted Hardin's account of constitutions as conventions (and conventions as coordination solutions), it has responded to this fact not by adopting the theory but by ignoring it.[4] Hardin's work has few partisans. At least as strikingly, it has few worthy opponents; it is not considered among the main accounts of mass democracy that requires mention and refutation. Why this relative non-reception?

The first part of my tentative answer is this: in a field divided between democratic theorists, liberal theorists, and historians of political thought, convention theory embarrasses all three by seeing and raising their characteristic claims and points of pride. It captures too much about real democracy to sit well with self-styled partisans of "Real Democracy." It is too accurate regarding the real preconditions of diversity and choice to sit well with those who imagine those values are best furthered by a universally valid rational consensus. It is too successful at explicating Madison's constitutionalism, and Hume's, to sit well with those determined to mine the

history of thought for lost doctrines of virtue and the common good, and to slough off insights that glitter less. In other words, convention theory is not just counterintuitive but *embarrassing*: it accomplishes too much without the aid of premises regarded as necessary by too many.

If this first part of the answer involves an amateur sociology of knowledge, the second part is more substantive. Hardin's formulation of convention theory sometimes makes it appear more pessimistic and quietist than it need be. Conventions may appear to lock in inequality or bias, in the sense of providing universal benefits but doing so unequally. And they may appear to achieve stability at the cost of a sort of rolling hostage-taking: the status quo is to be acquiesced in because, and for at least some parties only because, the costs of "re-coordination," including the risks of social strife, are too high to be worth it. But one of Hume's central claims is that political institutions and practices resemble scientific discoveries in being (lumpily) progressive: institutions and practices are human technologies resulting from trial, experience, and diffusion. I submit that certain political technologies, largely unknown to Hume and insufficiently stressed by the partisans of convention theory, go a long way to addressing the theory's critics—both those whose criticisms are explicit and those whose critique takes the form of scholarly neglect.

Theory Out of School(s)

Democratic Theory

A great many contemporary political theorists define themselves as "democratic theorists." One would think this would give them a profound interest in studying a theory, like that of convention, which explains the nature and justification of democratic institutions. But in using the same word, democratic theorists do not always mean the same thing.

To dramatize the point, consider two paintings (not, alas, reproducible here, but available online).[5] They were painted in the same year (1943), share a more or less epic style, and even share a title: *The Four Freedoms* (after Franklin Roosevelt's famous speech on that theme). But they exemplify two profoundly different approaches to democracy: one stirring but hardly relevant to modern problems; the other, much more pertinent but alas, less inspiring.

Norman Rockwell's painting, "Freedom of Speech," is very familiar, as is its message. A plainly dressed man rises to speak at a town meeting, with a

dignity and presence that make others, whose dress and accessories mark them as more learned and richer, pay attention. The painting, along with Rockwell's paintings on the other three Freedoms, is the graphic version of a civics book, portraying homespun emotions and ordinary citizens. (The other paintings portray a Thanksgiving table groaning with food [Freedom from Want]; a multicultural group of worshippers, each engaging in his or her respective prayer [Freedom of Worship]; and parents checking on their children, safely in bed, while holding a newspaper chronicling a bombing elsewhere [Freedom from Fear].) The other work, by Hugo Gellert, is much less famous than Rockwell's, as well as more realistic, more political, and less comforting. Gellert's painting portrays organized power. Freedom from Worship—the atheist's Geller's least interesting panel—portrays abstract symbols, not particular worshippers: this is freedom of *organized* religion. Freedom from Want portrays stylized foods, not citizens sitting down to food. Freedom from Fear is represented not by childish innocence that lacks the *feeling* of fear, as in Rockwell, but by an imagined defeat of the *cause* of fear: a Nazi snake being crushed by a strong arm.

Most fascinating for current purposes: Gellert's "Freedom of Speech" is symbolized by a microphone and a book—a blank book, indicating that the point is not what is said but the speaker's power to say it and be heard. That is, the real power of free speech attaches not to every anonymous speaker at a town meeting—who may indeed speak but will be heard and heeded by very few—but to the person who commands a microphone and the prominence to speak to a mass audience either in an auditorium or, more likely, over the air (one imagines a Roosevelt Fireside Chat).

The panoply of freedoms is presided over by a huge figure of Franklin Roosevelt, portrayed with such realistic toughness as to look slightly menacing. The message is clear, and clearly one of coordination: to save our freedoms in a time of peril, we must get with the program and follow along with some authority, leader, or focal point that others can be counted on also to follow.

Leadership is not the only way of solving coordination problems. A liberal or a non-authoritarian democrat should shrink—as the pro-communist Gellert did not—from the implications of relying on leadership on a mass scale, of resting freedom, and social values generally, on the prerogative of one national politician.[6] Of the standard means of solving coordination problems—focal points, leadership, common knowledge, and conventions—only the last is fully and reliably consistent with both the

need for concerted action and the imperative that the exercise of power be limited (Sabl 2012).

The illustration on the cover of Hardin's *Liberalism, Constitutionalism, and Democracy* portrays the framing of the US Constitution. But while the face-to-face coordination that founded the constitution makes for a better graphic, the day-to-day coordination whereby every citizen more or less unthinkingly and automatically acquiesces in established constitutional authority is, for political purposes, more significant.[7] There is nothing wrong with calling a book that treats liberal democracy as a matter of large-scale conventions *Liberalism, Constitutionalism, and Democracy*. But a simpler title might have been *Real Democracy*.

The actual recent book called *Real Democracy*, by Frank M. Bryan, places on its cover neither constitutional framing, nor everyday acquiescence, nor populist leadership, but precisely Rockwell's *Freedom of Speech*—and the "real" democracy it has in mind is the town meeting. In general, most recent democratic theorists, participatory or deliberative in inspiration, would consider town meeting democracy more "real" than the variety through which we choose representatives and they settle legislative matters by majority vote. What lies behind this?

One version of participatory democracy prizes the agency involved in small-scale self-government, where each ordinary citizen can hope to make a difference through speaking. If one is determined to value democracy in this sense and for this reason, it will follow that democracy can be studied most fully where the number of participants is small. As a corollary, democratic theory will have little to say about—indeed, will be very tempted to discount and demean—decisions on the scale of a modern state, and will disparage the study of constitutional structures that promise to make democratic politics more secure and durable while retaining some popular influence on collective decisions.[8]

More interesting for current purposes, because less obviously determined to ignore the most salient and inescapable facts about modern society, is Josiah Ober's attempt to explain Athenian-style participatory democracy as a regime of *coordination* that made only limited and informal use of constitutional methods (Ober 2008). Ober portrays a world of common knowledge and common purposes. While he acknowledges in various places that modern mass polities does not possess these qualities, he still implies that we can draw more analogies between Athens' situation and our own than is credibly the case.

Ober does not claim that Athens' scale (about 250,000 residents, of whom tens of thousands were free male citizens) allowed each individual to make a profound difference: the numbers involved were too large. But he is determined to argue that various aspects of Athens' design—from the inward-facing structure of its assembly to the division of citizens into heterogeneous tribes—allowed it to operate through *common knowledge* solutions to coordination problems, especially the problem of ensuring that each citizen could know that others would do their part in governmental and military affairs. In fact, Ober mentions *only* common knowledge, and not mass leadership or conventions of authority, as the known means by which citizens can coordinate (Ober 2008: 114). (For obvious reasons, common knowledge would be, for participatory democrats, preferred over authority, constitutionalism, or leadership *as* a means.)

Ober regards Athens as diverse in many ways, including with respect to specialized knowledge stemming from economic roles. (He cites evidence that Athenians recognized at least 170 different occupations (Ober 2008: 21n34)—though one might observe that the US Bureau of Labor Statistics currently counts nearly 8500, including Annealing Furnace Operator, Chum Driller, Computer Numerically Controlled Shot Peening Operator, and, no joke, Pepper Picker).[9] Ober insists, however, on a striking lack of diversity in one sense: Athenians, on his view, possessed "shared core preferences" that included the *polis's* prosperity and military strength in a context of deadly inter-polis competition.

The assumption of shared values is absolutely necessary if coordination problems are to be solved by common knowledge alone. For one thing, it rules out rent-seeking; in another piece, Ober argues that Pericles' leadership posed no danger to Athens and came at no cost to democratic agency since

> there was no contradiction between seeking his individual good and the good of a powerful and flourishing community. As Athens flourished, so too did Pericles. (Ober 2006: 151)[10]

More generally, the assumption of shared values is necessary because in order for common knowledge to work as a method of coordination, the individuals who share knowledge *must broadly want the same thing*. If my purposes clash with yours more thoroughly or more fundamentally than they coincide, the knowledge that we have in common will consist of the fact that you and I must be enemies in a fairly literal sense. As Hardin has

noted, we may then need to occupy separate polities, or a single one that suppresses diversity through authoritarian means (Hardin 1999: Chaps. 5 and 7).

Aware that modern polities lack this level of shared values, Ober admits that Athenian democracy resembles modern states less than it does modern *firms*, whose members have a common interest in firm survival and profitability on an impersonal but limited scale and in a context of fierce competition (Ober 2008: 90 and passim). But one may draw the further normative conclusion that participatory democracy is normatively attractive *if and only if one thinks it possible and desirable to have a civic culture at least as strong, coherent, and action-focusing as a corporate culture*. And this is likely to seem possible and desirable in turn if, and only if, one regards democratic life as a matter of constant crises that the polity must face as a whole (as opposed to its citizens facing various problems severally). There are some indications that Ober does see democracy this way, especially with regard to environmental crises (Ober 2008: 5). More generally, theorists of participatory democracy quite commonly fault citizens for ignoring crises that seem obvious (to the theorists) (Wolin 1969).

The question of how polities are to coordinate their decisions is therefore a crucial one. One's answer to the question clarifies not only what one means by democracy but what one means by choice and political agency (as well as what one means by coordination, an issue discussed below). Participatory democracy, even in its most sophisticated versions, envisions a polity acting as a single citizen body at the cost of denying individual citizens the prerogative of choosing their own purposes and dissenting from the prevailing sense of common interests. In contrast, conventions of authority are not themselves matters of agency, either as subject or as objects. The convention does not collectively "do" anything in particular (only designating how decision makers will be chosen), nor do citizens choose it (each citizen mostly acquiesces in the convention that exists, rather than being able to form a new one). But there seems a systematic sense in which conventions allow much greater agency in the realm of personal judgment, including judgments about politics, than is conceivable under the common knowledge that undergirds participatory democracy. Participatory democrats typically argue that polities must engage the citizen body, acting as one, in order to address crises. I submit it is largely the other way around: they are determined to see politics as a rolling crisis because that allows them to discount the costs of a politics that prefers a feeling of common agency to a diversity of individual purposes.

Liberal Theory

Proponents of *liberal* theory are aware of this problem and aim to avoid it. They value diversity of life plans and of "conceptions of the good" (the systematic, coherent accounts of how to live that John Rawls imagines most persons have the capacity to formulate) (Rawls 1999). I have mentioned that convention theory has no need for normative consensus; as long as people accept the direction of the same set of authoritative conventions, they need not agree on, or even think about, why; in fact, their reasons are likely to differ radically. Hardin notes the reality-based advantages of seeing politics this way: while there are a great many social theorists who have claimed that shared values are needed to maintain a stable polity, fewer (i.e. none) can list, with good evidence, what those shared values are (Hardin 1999: 9, 280 and elsewhere).

One central reason that prevalent liberal theories posit the need for normative consensus is, of course, that doing so seems to allow for greater *critical purchase* than a theory that explains social order as the habitual acceptance of conventions. The fear is that unjust orders, or those that discount the rights and interests of certain classes of citizens, might gain the ready acceptance of the advantaged and the resigned acceptance of the disadvantaged who doubt that great change is possible. A "proper" liberal theory, i.e. one that distinguishes between political structures that are normatively justified and those that are not, is supposed to allow only the kinds of diversity and choice consistent with justice and equality, as suitably defined in turn by that theory. A customary order that cannot justify itself is to yield to a more reasonable order that can.

Quite often, the argument rests on the explicit or implicit premise that so-called ideal theory—systematic accounts of justifiable social orders—is a precondition for social change. The (almost always implicit) premise is that ordinary people will not seek social change if all they have are urgent felt grievances and rough-and-ready, easily understandable accounts of what might be done to address those grievances. Absent such, they will wait, helpless and hopeless, for an abstruse book to provide them with proper goals. To the extent that this premise is—palpably—false, the need for critical theory is diminished. But often the critical demands of liberal theory are articulated, more credibly, in normative terms: even if a social movement or reform campaign *could* arise without the help of liberal theory, it would not *deserve to succeed* unless it could pass the tests proposed by that theory.

The striving, in other words, is not for social criticism or reform but for *rational* or *justifiable* social criticism and reform.

Here convention theories of liberalism again seem both superior to rationalist theories and acutely embarrassing to them, on two grounds. First, there is a problem of justificatory circularity. To the extent that ideal theory is supposed to provide critical purchase, so that discovering and defending the right theory will bring about more justice or equality in the world, it would seem to be necessary that the theory stick fairly closely to principles and modes of argument that ordinary people in fact recognize (along the lines of Michael Walzer's "connected" criticism[11]). But the more a theory aspires to the rational consistency, technical specificity, and comprehensive scope of a moral-political system, the less likely it is to track folk beliefs. Such beliefs are, everywhere and always, incompletely theorized, piecemeal, and partial (in every sense). Most saliently, ordinary citizens put forth their social opinions in the expectation that they will be accepted as they are—on the axiomatic assumption that democracy means each being entitled to his or her own opinions—*rather than* put to a test of rational and intersubjective acceptability. To express a similar argument in empirical rather than a priori terms: if the goal of *A Theory of Justice* was to bring about a society whose economy observes the difference principle, it has not, as Raymond Geuss has observed, worked very well (Geuss 2005).

Put more simply: convention theory suggests that ideal theory is neither necessary for either constitutional stability or needed change, nor even particularly helpful to either end. It implies that rational justification of entire social orders is both a fairly eccentric desire and an almost completely dispensable one. By making these suggestions, convention theory guaranteed its enforced obscurity. A theory that disagrees with one's answer to one's favorite questions is to be refuted; but a theory that mocks the importance of the questions can only be ostracized.

A second way in which convention theory flouts not only the premises of high liberal theory but also (and worse) its sensibilities has to do once again with common knowledge. Rawls regards the principles of justice as a stabilizing force because they do, or would, embody common knowledge: the parties in the original position are choosing principles of justice under the assumption that there will be "general awareness of their universal acceptance."[12] This seems a wildly implausible account of how any liberal society could operate. We would all have to be monitoring one another constantly, from birth and unto death, not for external behaviors but for inner states of mind. (Compare the Woody Allen joke that he once cheated

on a metaphysics exam by peeking into the soul of the kid sitting next to him.) One overheated critique of *A Theory of Justice* claimed that it rested on a quasi-totalitarian plan of education and indoctrination (Schaefer 1979). No serious critic thinks that Rawls in fact had such in mind. It is more accurate to say that Rawls hoped that political activity and debate within liberal institutions would educate citizens in liberal principles through a sort of normative invisible hand.[13] But one important reason that Rawls and his followers do *not* think that they need indoctrination is that they fail to acknowledge the Madisonian, conventionalist point that a free society produces not only a huge variety of social and moral opinions but also widespread ignorance in one quarter of what is believed in others. In actual democracies, people regard opinions deriving from unfamiliar circumstances, backgrounds and situations as both surprising and presumptively illegitimate. The folk definition of "special interest" is an interest held by geographically and/or culturally distant sorts of people with unfamiliar concerns. Most Nebraskans think that mass transit is a special interest; most New Yorkers think the same of farm subsidies (Hibbing and Theiss-Morse 2002).

Hardin's claim that Humean convention theory is exclusively positive, and only normative or prescriptive occasionally and when Hume gets overexcited (Hardin 2007: passim), to my mind goes too far. But it makes sense in the context of ideal theory and as a reaction to it. It is very common for high liberal theorists to regard their theorizing as a political act, designed either to enable social criticism or to buttress a new and better social order.[14] Hardin's position reflects the accurate conviction that this hope, in its usual form, is mostly vain. No theory complex and counterintuitive enough to be innovative and intellectually challenging will be sufficiently simple and uncontroversial to serve as common civic currency. Nor would we necessarily welcome the flattening of diverse private experiments of living that would, in imagination, make it possible for any such currency to serve as moral tender. Hume himself thought that one of the few ways of calling an act contrary to reason was to show that it rested on false factual beliefs. And since he doubted that passions were (aside from the possible case of a few sages) directly subject to rational control, he suggested that the only way to reform one's behavior through philosophy might be by changing one's social commitments or situations so that the sentiments one valued would be likely to arise through new habits and experiences (Hume 1740, 1987: 168ff). To the extent that convention theory does allow for political and social reform, it must probably do so through versions

of those two insights. Both seem more authentically liberal than the wish for a society based on constant, low-level, moral harangues.

Philadelphia Versus Cambridge

The historical or contextual school of political thought—Cambridge School for short—has two goals: one explicit, the other universally understood but never written down. The explicit goal is to improve the quality of the history of ideas, a.k.a. intellectual history, by ruling out "presentism"—the impulse to make sense of past theories by interpreting them in light of current concerns. Instead, scholars are to master the language and the intellectual arguments current at the time an author was writing. We are to assume (reasonably) that authors were writing for an audience that understood language as it was used then, not as it is used now; and (more controversially) that their main goal was to intervene in the debates of their time rather than to effect a timeless contribution to knowledge.[15] The implicit premise of the *main factions* of Cambridge is simpler: liberalism is bad. In one version of the thesis, associated with Quentin Skinner and Philip Pettit, liberalism entails mere non-interference with individual choice, while republicanism involves "non-domination": the pursuit of institutions or practices that render it *impossible* for those with power to act arbitrarily against those who lack it (Pettit 1997). This contrast, somewhat ironically, makes the most sense in a specific but odd context: the British, in which a political class universally affirms liberal values while proclaiming the impossibility, or at least the badness, of constitutional checks on parliamentary sovereignty. In such a context, "liberalism" indeed seems like a "trust-us" ideology in which nothing but the ruling class's good will and good sense prevent government oppression. But the non-interference/non-domination distinction makes less sense in the rest of the world, where liberalism and constitutionalism are considered near-synonyms rather than opposites.

The contrast can only be saved by arguing for a particular *account* of non-domination. This account requires citizens to place an active commitment to republican government ahead of their private concerns for achievement or advancement, and claims that citizens can only vindicate their civil rights, and protect their private purposes, if they possess, and foster in one another, a common civic virtue and vigilance. This account indeed contradicts traditional constitutional or Madisonian liberalism. The latter form of liberalism assumes that politics will be of surpassing interest only to a few. It regards the main safeguard for liberty as a set of institutions that

harness ambition to check ambition and channel citizens' overwhelmingly private and potentially conflictual passions into pursuits that are harmless (peaceful religious competition) or beneficial (economic, artistic, or scientific emulation).

A second strand of Cambridge, sometimes called "civic humanist" and associated with J.G.A. Pocock, goes yet further. It insists that civic virtue is not just instrumentally but intrinsically good, that the highest human capacities can only be displayed in political action, and that a system based on the management of interests can only be an exercise in delay or evasion by a polity that has culpably flouted the imperative to seek a common good. On this view, the durability of a Madisonian system only compounds its errors, habituating a benighted citizenry to the false belief that politics exists to protect their diverse private purposes (Pocock 1975).

Convention theory challenges Cambridge on all these points—in ways destined, alas, to maximally embarrass it. While more than willing to admit that the specific, local terms in which a theory is expressed probably derive from intellectual context, it is not especially *interested* in those specific and local elements of a theory. A theory is interesting to the extent that it discovers something of lasting truth and value. Knowledge of contexts, on this view, may yield only tragic or negative lessons, as when Hardin argues that the central insights of Hume's convention theory were neglected for 200 years because they flouted the ethical frameworks of his time and could only become manifest after the re-discovery of coordination, expressed largely in mathematical terms, in the twentieth century (Hardin 2007: 25–26).

When it comes to civic republicanism or civic humanism, the particular brand of liberalism represented by convention theory not only denies the need for a common standard of civic virtue, but takes great pride in doing so; stresses the advantages of doing so; and implicitly mocks republican and humanist theories for even attempting to seek common values (as mentioned above). Conventions are, on the contrary, to be prized for allowing us to enjoy a common political and social order in the midst of a degree of social and moral diversity that renders impossible agreement—even "overlapping" agreement—on what the moral basis of that order might be. (Nor does it much matter, since the order does not rest on moral agreement.) And one notable tendency of convention theorists is to celebrate the fact that actually existing liberalism allows for this moral diversity.

Again, convention theory is objectionable not because it differs with Cambridge answers but because it mocks Cambridge questions. Whereas "high" or "ideal" liberalism often meets critics of liberalism halfway, assuring them that liberalism, too, aspires to a normative consensus and a respect for community values that are not too different from republican (or, in an earlier version, communitarian) virtue,[16] the liberalism of convention refuses to grant even half-hearted affirmation to those who believe that liberty requires more than a modicum of civic virtue, who assume that the kind of conformity characteristic of small towns is something to be admired rather than escaped, or who insist that their fellow citizens must persistently prefer public business over their often exciting and deeply fulfilling private projects.

Conventional Wisdom and Political Learning

As said, while convention theory's relative lack of take-up can be explained in large part by its offense to prevailing prejudices, it is also the case that convention theory has not always put its best foot forward. It has not always sufficiently addressed potential weaknesses, even when these are only apparent or contingent.

Biased Solutions and Accreted Powers

The first set of concerns derives from the acknowledged fact that political conventions of authority, a.k.a. constitutional or fundamental conventions, are not pure coordination games in which the parties are indifferent regarding the solutions provided that one solution is durably reached. (Which side of the street to drive on is the most commonly mentioned real-world example of a pure coordination problem. It may in fact be the only one: solving *any other* coordination problem creates relative winners and losers.) Like most coordination problems, constitutional conventions reflect an "impure," "biased," or "bargaining" problem in which all parties have a clear interest in reaching some solution rather than none, but each benefits most from certain solutions and is relatively disadvantaged by others.

One is tempted to say—and this is one of many instances in which Hardin underplayed his own potential theses—that *only impure or biased coordination games are politically interesting and potentially permanent, precisely because pure ones are easily solved through strategies of mutual observation and communication*. Only when a solution benefits everyone

greatly but different groups unequally is it likely to preserve the problem: such solutions are stable enough to preserve the conditions of politics (i.e. relative stability, not constant war or anarchy) yet controversial enough to remain political issues, matters of contestation. To focus one's discussion on pure coordination games bespeaks a strong desire to see polities as overwhelmingly consensual, as seeking methods of making sure that all find ways to coalesce in acting for the common purposes they would like to pursue together. (At the extreme, some theorists of common knowledge refer to coordination games when they really mean *assurance* games, in which a single solution is preferred by all if only each can be reassured that the others will find it too.[17]) Impure coordination games are the stuff, on the contrary, of real politics: citizens' interests and their purposes are different and partly conflicting.

The matter is in fact not so simple, since the success of a constitutional order requires, as Hardin aptly notes, a good deal of uncertainty regarding which specific parties a set of institutions will benefit in the future (Hardin 1999: 129–134). Still, constitutional decisions inevitably render some society-wide arrangements more difficult than others.[18]

As I have argued elsewhere (Sabl 2012: Chaps. 6 and 7), there are really two kinds of bias involved. One might be called "vertical inequality," and may be seen as a writ-large version of rent-seeking, though its importance transcends the connotations of that label. Those whom a constitutional system designates as having the authority of office may use their unique prerogative over authoritative decisions to grab resources, powers, or immunities for themselves or their associates. A particularly egregious example of this is that governments can leverage the authority deriving from their unique ability to provide public goods in order to engage in acts that formally resemble public goods (e.g. defense) but may promote the glory of leaders rather than the good of citizens (e.g. wars of choice). Hardin certainly notes this possibility, but does not stress it. A second kind of bias involves "horizontal" inequality: the ability of groups that win power to give disproportionate rights and opportunities to one sector of society, perhaps the majority, at the expense of others. (To see the distinction between vertical and horizontal inequality: an African-American of an impoverished background who is elected mayor of a city may be able to name any number of African-Americans to city office—but unable seriously to affect housing segregation or the differential quality of mostly-white and mostly-black schools.)

Hardin's treatment of convention acknowledges both kinds of inequality. The first he treats repeatedly but rather briefly. However, those who regard war-fighting and empire as the most salient and nefarious activities of the American state (e.g. such political theorists as Sheldon Wolin and George Kateb, ever scarred by the experience of the Vietnam War) might wish the treatment were fuller.[19] Hardin's treatment of what I call horizontal inequality, under the name of "unequal coordination," is more extended but not particularly optimistic. Hardin argues that given "successful coordination on a constitutional regime that does not seem to give equal standing to some group," "[t]hat group's members may nevertheless find their interest is to acquiesce in the coordination. . .. The price of mutiny may be too high for any benefits it might bring" (Hardin 1999: 306). Hardin explicitly mentions "the ghetto poor" in this context, and in an Appendix seems to imply that while universal welfare schemes might be able to latch themselves onto existing coordination schemes, explicitly redistributive, reparative, or anti-welfare schemes do not serve the advantage of whites and are therefore, on a coordination view, more or less out of luck (Hardin 1999: 307, 328–331).

The usual remedy for vertical inequality in the form of rent-seeking is a separation of formal powers, along with legal and social protections for civil-society institutions with a professional interest in checking power. Given a professional interest in investigating corruption and durable protection for the activity of doing so, legislatures, courts, and various activists and journalists in civil society can be counted on to do so (without the need to assume a uniform "civic virtue") (Philp 2007). Such methods are probably insufficient to prevent mischief in the arena of foreign warfare: the secrecy necessary to meet genuine national security threats also prevents the oversight needed to expose fraudulent ones (Sagar 2014). But if convention theory provides no clear solution on these matters, neither does any other account of government. At least seeing foreign glory as a form of rent-seeking, stemming from legitimate and probably unavoidable coordination-based power, might help clarify and sharpen the problem.

When it comes to horizontal inequality, such a *tu quoque* argument might also work. (Not everyone who faults others' accounts of politics for not addressing racism and poverty has a politically plausible roadmap for ending racism and poverty.) But we can do better. Consider Adam Przeworski's well-known minimalist or "paper stones" defense of democracy, in which voting represents a peaceful signal of the social power that social forces could potentially muster in violent fashion (but do not need to,

because they express their power by voting instead). Less often noted is that Przeworski develops this account with explicit reference to Hardin's (early) account of constitutionalism as stable because re-coordination, or "mutiny," would have costs that are too high (Przeworski 1999: 46–47). Przeworski does not really develop the implications of his account in terms of convention theory. One can easily do so, however, provided that one avoids a reductively quantitative account of social power. Voting works, on Przeworski's view, because those tempted to mutiny will refrain from doing so if the voting process reveals them to be a minority. This may be true in a very uncompromising sense of mutiny, perhaps the Marxist sense. Small minorities with no particular social or military power rarely win civil wars and take over a society's monopoly on force; there is a reason Marxists believed, and had to believe, the proletariat and its allies to be an "immense majority."

But mutiny and re-coordination do not need to be about winning. *Unsuccessful* rebellions—which are usually, being unsuccessful, called riots—impose substantial costs on the victors. If a group competes in mass elections and wins a substantial chunk of the vote, though not a majority, it obtains a strong *and cheap* signal, to itself and outsiders, that its demands could become the object of costly, though losing, civil unrest.[20] Electoral and legislative contestation thereby serves as an important check on horizontal inequality; by demonstrating the possibility of a costly mutiny, it may scare the powerful into considering re-coordination. And to the extent that legislative bodies represent a plurality of social voices more effectively than a single executive can, the fact that the latter must concur with the political decisions of the former represents a limit to the power accruing to coordination.

Institutional Reform Within Formal Continuity

Convention theory appears to explain why political institutions persist, but not how they change. To be sure, the desire for constant institutional innovation for its own sake, evident in Dewey's work and more recently Archon Fung's, is almost certainly a minority taste (Fung 2012). A great many people find radical change even in their consumer goods quite disorienting. A computer-run car with an electric motor could *theoretically* mount its controls in any way one likes. But there is a reason that Priuses still use steering wheels and pedals for steering, acceleration and braking, even though the mechanical reasons for such control devices no longer obtain. This is all the more valid for political institutions, which must provide not

only the comfort that facilitates occasional operation but the continuous and absolutely necessary stability necessary to plan the rest of life. Still, it is common, and legitimate, for those who live under constitutional regimes to want politicians to be able to tinker and adjust them to meet new needs. It might appear that convention accounts of coordination leave little room for doing that. Every substantial change seems to risk misunderstanding, non-coordination, the re-opening (or new opening) of disputes that productive stability requires us to set aside.

Convention theory has an answer to this, but not an answer that it always sufficiently stresses: *de facto institutional change and development within formal conventional continuity*. Critics of American constitutionalism—and in the international arena, there are many more critics than fans—often note, correctly, that the American constitution is in international comparison extremely hard to amend[21] and contains very few specific guarantees or stipulations that the system must benefit ordinary citizens in tangible ways (by providing them health care, or free education, or jobs). But the second flaw, if such it be, compensates for the first. Precisely because there are no specific guarantees, our system has evolved a norm of re-interpreting—in what might seem implausible ways—extraordinarily vague terms like "equal protection," "general welfare," and "commerce" so that the system will in fact serve purposes whose achievement benefits more or less everyone but whose pursuit seems disallowed by the formal rules. Hardin stressed the degree to which the constitution at its origins had commerce as its central purpose, in opposition to the anti-Federalists' goal of resting the new republic on local farming and the civic virtues that they quaintly believed unique to that. One could equally note, however, that the purpose of "commerce" has since been stretched to include the comprehensive regulation required for the flourishing of industrial and post-industrial societies. Similarly, "the executive power," once exercised by a President and by a few aides (called "secretaries" in the US system for the good reason that they once handled their own correspondence and wrote their own reports), has been stretched, without *formal* breach of convention, into a full-scale administrative state. In other words, re-coordination by stealth renders it more possible than might at first appear to combine the advantages of relative constitutional stability and those of necessary institutional change.

Technologies of Re-coordination

The concern with horizontal inequality may be restated, and amplified, by taking the perspective of those who are planning strategy on behalf of relative losers. One of the central insights of *Liberalism, Constitutionalism and Democracy* is also the one most hotly resisted by students when I teach it: that it could be in the interest of parties *relatively* disadvantaged by a scheme of coordination to acquiesce to it indefinitely. After all, they do benefit from the existing scheme to a very considerable extent compared to having no constitutional scheme at all; they might not be able to secure the acquiescence of the more privileged in a project of "re-coordination" under a different and perhaps fairer scheme; and the costs of seeking re-coordination, likely to be paid in civil strife, may be very high, especially for the disadvantaged themselves.

In a sense this argument is very strong, precisely because of the disanalogy between constitutional conventions and other instances of bargaining and unequal gains. If workers prefer having a job to not having it yet feel their wages and working conditions are unfair, and lower than ones that the employer can reasonably grant, they can strike—forgoing mutual gains from the employment contract, but only temporarily and in the service of a new contract on better terms. But this only works because the rest of society goes on peacefully during the strike: conventions of authority abide. A strike against conventions of authority themselves looks like anarchy or insurrection. Hume says in this context, more or less, that disputes over property can be settled by authority, but disputes over authority are only settled by "the swords of the soldiery" (Hume 1740: 3.2.10.15).

As in many cases, however, new political technologies can expand these alternatives. One of these is free speech, which in its more radical forms was too free for Hume's taste in his day, but still, with respect to fundamental constitutional questions, very constrained compared to what we are used to now. (The Seditious Meetings Act two decades after Hume's death illustrated this: British elites were not willing to allow free discussion by those who approved of the French Revolution, even if they refrained from drawing explicit British lessons.) Another, more to the point, is civil disobedience. Civil disobedience is a complex and theoretically disputed act—which is actually something of a problem; it would send a clearer signal if governed, like war, by rough norms, so that people could know when a violation of expectations was intended. Still, its core and classic meaning, as expressed in the movements led by Gandhi and King, was to *reject the legitimacy of prevailing institutions* while expressing *willingness, shown by*

lack of personal violence, to live peaceably under a different arrangement of authority and civic status.[22] Civil disobedience, in this way, promises radically to reduce the costs of attempting re-coordination, and thereby to make it more attractive. A corollary of this is that civil disobedience, as compared to violent rebellion, makes it harder for the privileged to insist on the status quo as the only alternative to constitutional indeterminacy and palpable anarchy. Civil disobedience, so conceptualized, remains a very radical strategy and one that threatens to backfire if its practitioners' claims to justice cannot be vindicated or the terms on which they will agree to acquiescence are not clear.[23] Still, it represents a major innovation with respect to political action and a crucial friendly amendment to convention theory.

Conclusion

I have suggested that the convention model of liberal democracy has faced mostly unfair and irrelevant criticisms. These have, in turn, obscured some other criticisms that are more pertinent but can be answered. I would conclude by stressing that these answers address not only rational worries but also what might be called emotional or rhetorical ones. That is to say: we demand of theories that they give us not just truth, but usable truth. While the goal of political science, like that of all sciences, is to increase our knowledge, the *point* of political science, like that of all sciences, is to increase our ability not only to make sense of the world but to have a sort of purchase on it, an awareness of how it might be made somewhat better through interventions in the causal process. (This is to put things in the least utopian terms possible.) I fear that convention theory can be seen as not doing this. It seems to be telling us to prize our existing institutions for fostering wealth creation and a multitude of personal projects, without giving us a sense of how political institutions themselves might be improved. As Michael Freeden has noted, there is an odd disconnect between today's self-styled liberal political theories, which stress equilibrium, stability, and a tendency to stick like glue to static principles of justice or equality if they are ever achieved, and a longer liberal tradition that stresses a "zeal, eagerness, insistence" for change and a fervent impulse for reform (Freeden 2005: Chap 1). Edmund Fawcett's recent popular account of liberalism has stressed this dynamism to such a degree that liberalism is said to favor *no* permanent solutions: only certain values to be vindicated in an awareness that constant social change requires constant political adaptation (Fawcett 2014).

Convention theory, with its stress on the need for political and institutional stability, might seem allied with the former approach to liberalism. I propose interpreting it in a way that gives it much more in common with the latter. Because it doubts that institutions reflect, or can reflect, eternal moral principles grounded in reason, it can recognize the utility and authority of institutional change as new social interests manifest themselves or new ways of satisfying old interests prove their worth. (On this matter, the wise can exploit the foolish. If a country, state, or province pursues a policy for ideological reasons, in advance of empirical evidence that it might work, other, more prudent political units can observe the experience of the imprudent innovators and draw lessons from that experience, learning "the easy way" from others' failures and successes.[24]) And while convention theory doubts the wisdom of seeking revolutionary change, it has plenty of room for steady and, over time, profound evolution within a framework of formal stability. It is true that there are intractable disputes (slavery, existential distrust among groups like the Hutu and Tutsi) that liberal-democratic institutions cannot address, as well as many private and technical goals whose achievement liberal-democratic institutions allow for but do not themselves bring about. But despite these caveats about what convention-based accounts of politics do not promise, it is crucial to acknowledge the progress that they do allow, encourage, and welcome.

Appendix: Two Models of "Coordination"

1. Assurance

6 6	0 4
4 0	4 4

Fig. 1 "Coordination problem" qua assurance (Re-drawn from Chwe (2001: 102); compare Ober (2008))
Solvable by common knowledge and real-time monitoring/signaling

2. Impure/Biased/Bargaining

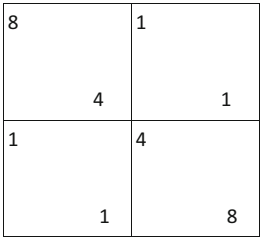


Fig. 2 "Coordination problem" qua impure/biased/bargaining (Hardin, translated from the ordinal; Schelling (1960))
Due to partly conflicting interests, not solvable through common knowledge plus monitoring because these would yield no determinate solution.
Only personal or constitutional authority solves (temporarily and subject to contestation).

Notes

1. The politics in question consisted, roughly, of (democratic) socialism cum participatory democracy.
2. One reason the theory has found it hard to gain purchase might that the relevant volumes embody diverse genres. Hardin (1982) is mostly formal, though non-technical. Hardin (1995) is more or less a work in comparative politics. Hardin (1999) is largely historical. And Hardin (2007) is interpretive. To understand the theory's full implications and force, one must, alas, read all four.
3. Compare Sabl (2012), Chapters 1 and 4.
4. An anecdote: when I mentioned to a top Yale graduate student that Hardin's work was underappreciated, he replied, somewhat surprised, that *One for All* was quite widely assigned in courses on ethnic conflict. When I clarified that I meant the work was underappreciated in political theory, he replied, "Oh, *of course* it's not read in political *theory*."
5. Norman Rockwell's *Freedom of Speech* may be viewed at https://www.nrm.org/2012/01/normanrockwells-four-freedoms/#post/0; Hugo Gellert's *The Four Freedoms*, at http://collection.whitney.org/object/43447.
6. The power attaching to leaders who are in a position to solve coordination problems, as well as collectively to a government granted this power by institutions, is noted by Hardin (1999: 102, 107, and esp. 114) and appears as early as Schelling (1960, Chap. 3). For an extended treatment see Calvert (1992). Ian Kershaw's short book *Hitler* (2000) could be seen as a primer on

how to leverage the coordination power attaching to the extra-constitutional role of leader, or *Führer*, into vast and terrifying power. One might note that the Nazi term for its totalitarian policy of requiring all civil society and voluntary groups to align themselves with Nazi ideology was *Gleichschaltung*, whose literal meaning is ensuring that all railways use the same gauge of track. Nazification was conceptualized, in other words, as a coordination problem. Hitler solved it.

7. The cover is available at https://global.oup.com/academic/product/liberalism-constitutionalism-and-democracy-9780199261680?cc=us&lang=en&; it is a detail, in mirror image, of a painting whose provenance I could not establish.
8. See the various writings of Sheldon Wolin (1994a, b, 2004).
9. http://www.bls.gov/soc/soc_2010_alphabetical_index.xls, accessed 21 October 2015; the latest update mentioned there is January 2013.
10. Compare *Democracy and Knowledge*, 11, where Ober's account of coordination omits the possibility of biased or impure games: all sink or swim together.
11. Walzer puts forth this idea in many works, notably *The Company of Critics* (1988).
12. Rawls (1999: 115, 115 n 8) cites David Lewis' philosophical account of convention as an explication of common knowledge.
13. On this, excellent is Bøyum (2013).
14. It is, however, uncommon in the Anglo-American world to *admit* to regarding moral philosophy as "one form of political action." An exception, containing that quotation, is Goodin (1988: ix).
15. An implicit minor premise, at least some of the time, is that political theory is a non-progressive discipline. Unlike science, on this view, it cannot hope to build on past discoveries (perhaps through the use of technical as opposed to everyday language) and document discoveries for use in the future.
16. For the most classic expression see Gutmann (1985).
17. Thus when Chwe (2001: 102) spells out in formal terms what his book has been calling "coordination" problems, the payoffs are those of an assurance game (see the Appendix). Ober cites Chwe at several points and in two places unmistakably describes an assurance game in prose: "If I know you all will fight, then I will too"; "Building common knowledge in public institutions addresses the 'carry through' problem faced by people with shared goals, but who will not individually act to achieve them unless each believes that others will act likewise" (Ober 2008: 179, 191; cf. 192, 194f, 199f.). Naturally, common knowledge and social monitoring solve *that* problem. They do not solve the ("impure," "bargaining") problem in which different groups in society share an interest in peace and order but benefit differentially from different forms of order. That requires authority, whether personal or constitutional, which will in turn give rise to disputes over authority.

18. For instance, the framers' decision to give commerce pride of place in the constitutional system made it unlikely that the system could accommodate large-scale land reform as the compensation for plantation slavery and the culmination of Reconstruction (Du Bois 1935).
19. Thus, Wolin in 1997, two years before *Liberalism, Constitutionalism, and Democracy*, describes postwar American elites as having produced "corruption, constitutional violations, incalculable death and destruction visited upon hapless populations abroad, steadily worsening racial relations, deepening class divisions, discreditation of the idea of public service (except for convicted felons) and, not least, a political system that large numbers of Americans wish to disown" (Wolin 1997: 154). This is not, to put it mildly, Hardin's most salient or dominant assessment of American constitutionalism.
20. There is also the possibility of partial disallegiance. In the 1980s in South Los Angeles, the police lost the trust of the local population to such an extent as to be considered merely a particularly well-organized gang. Such disallegiance has substantial costs to the larger community—again, partly expressed through potential or actual civil unrest—even though local residents made no effort towards formal revolution, towards founding a new city government or opting out of governmental institutions that seemed immediately useful, like roads and schools.
21. See, most recently, Tuck (2015).
22. The first of these claims may sound shocking to those who see the Civil Rights movement as broadly accepting American institutions. But I believe the historical case is very strong that Martin Luther King and other Civil Rights leaders endorsed the abstract principles that the Declaration and Constitution professed while denying absolutely that existing American society practiced them even approximately. See Lyons (1998) and Sabl (2001).
23. Thus Chong (1991) explains the decline of the Civil Rights movement after the mid-sixties partly as a consequence of the fact that its economic and social goals were vague, lacking a clear focal point or stopping point, once civil and political rights had been achieved.
24. In Sabl (2002), I call this the "paradox of innovation."

References

Bøyum, Steinar. 2013. Rawls's Notion of the Political Conception as Educator. *European Journal of Political Theory* 12: 136–152.

Calvert, Randall. 1992. Leadership and Its Basis in Problems of Social Coordination. *International Political Science Review* 13: 7–24.

Chong, Dennis. 1991. *Collective Action and the Civil Rights Movement*. Chicago: University of Chicago Press.
Chwe, Michael Suk-Young. 2001. *Rational Ritual: Culture, Coordination and Common Knowledge*. Princeton: Princeton University Press.
Du Bois, W.E.B. 1935. *Black Reconstruction*. New York: Harcourt, Brace and Company.
Fawcett, Edmund. 2014. *Liberalism: The Life of an Idea*. Princeton: Princeton University Press.
Freeden, Michael. 2005. *Liberal Languages*. Princeton: Princeton University Press.
Fung, Archon. 2012. Continuous Institutional Innovation and the Pragmatic Conception of Democracy. *Polity* 44: 609–624.
Geuss, Raymond. 2005. Neither History Nor Praxis. In *Outside Ethics*. Princeton: Princeton University Press.
Goodin, Robert. 1988. *Reasons for Welfare*. Princeton: Princeton University Press.
Gutmann, Amy. 1985. Communitarian Critics of Liberalism. *Philosophy and Public Affairs* 14: 308–322.
Hardin, Russell. 1982. *Collective Action*. Baltimore: Johns Hopkins University Press.
———. 1995. *One for All*. Princeton: Princeton University Press.
———. 1999. *Liberalism, Constitutionalism, and Democracy*. Oxford: Oxford University Press.
———. 2007. *David Hume: Moral and Political Theorist*. Oxford: Oxford University Press.
Harrington, Michael. 1983. *The Politics at God's Funeral: The Spiritual Crisis of Western Civilization*. New York: Holt, Rinehart, and Winston.
Hibbing, John, and Elizabeth Theiss-Morse. 2002. *Stealth Democracy: Americans' Beliefs About How Government Should Work*. Cambridge: Cambridge University Press.
Hume, David. 1740. *Treatise of Human Nature* 3.1.1.9.
———. 1987. The Sceptic. In *Essays Moral, Political, and Literary*, ed. Eugene F. Miller, revised ed. Indianapolis: Liberty Fund.
Kershaw, Ian. 2000. *Hitler*. London: Longman's.
Lyons, David. 1998. Moral Judgment, Historical Reality, and Civil Disobedience. *Philosophy and Public Affairs* 27: 31–49.
Ober, Josiah. 2006. Thucydides and the Invention of Political Science. In *Brill's Companion to Thucydides*, ed. Antonios Rengakos and Antonios Tsakmakis. Leiden: Brill.
———. 2008. *Democracy and Knowledge: Innovation and Learning in Classical Athens*. Princeton: Princeton University Press.
Pettit, Philip. 1997. *Republicanism*. Oxford: Clarendon Press.
Philp, Mark. 2007. *Political Conduct*. Cambridge: Harvard University Press.

Pocock, J.G.A. 1975. *The Machiavellian Moment*. Princeton: Princeton University Press.

Przeworski, Adam. 1999. Minimalist Conception of Democracy: A Defense. In *Democracy's Value*, ed. Ian Shapiro and Casiano Hacker-Cordón, 23–55. Cambridge, UK: Cambridge University Press.

Rawls, John. 1999. *A Theory of Justice*. revised ed. Cambridge: Harvard University Press.

Sabl, Andrew. 2001. Looking Forward to Justice: Rawlsian Civil Disobedience and Its Non-Rawlsian Lessons. *Journal of Political Philosophy* 9: 331–349.

———. 2002. When Bad Things Happen from Good People (and Vice-Versa): Hume's Political Ethics of Revolution. *Polity* 35: 73–92.

———. 2012. *Hume's Politics: Coordination and Crisis in the History of England*. Princeton: Princeton University Press.

Sagar, Rahul. 2014. *Secrets and Leaks: The Dilemma of State Secrecy*. Princeton: Princeton University Press.

Schaefer, David Lewis. 1979. *Justice or Tyranny? A Critique of John Rawls' A Theory of Justice*. Port Washington: Kennikat Press.

Schelling, Thomas. 1960. *Strategies of Conflict*. Cambridge: Harvard University Press.

Tuck, Richard. 2015. *The Sleeping Sovereign*. Cambridge: Cambridge University Press.

Walzer, Michael. 1988. *The Company of Critics*. New York: Basic Books.

Wolin, Sheldon. 1969. Political Theory as a Vocation. *American Political Science Review* 63: 1062–1082.

———. 1994a. Fugitive Democracy. *Constellations* 1: 11–25.

———. 1994b. Norm and Form: The Constitutionalizing of Democracy. In *Athenian Political Thought and the Reconstruction of American Democracy*, ed. J. Peter Euben, John Wallach, and Josiah Ober, 29–58. Ithaca: Cornell University Press.

———. 1997. The Destructive Sixties and Postmodern Conservatism. In *Reassessing the Sixties*, ed. Stephen Macedo. New York: Norton.

———. 2004. *Politics and Vision*. expanded ed. Princeton: Princeton University Press.

Collective Action in America Before 1787

Jon Elster

Russell Hardin's *Collective Action* (1982) is a classic of political science. In this chapter for a volume honoring him, I pursue some ways in which the American colonies and states, as well as the citizens of these polities, were subject to collective action problems, and occasionally and temporarily managed to overcome them. Although my ultimate motivation is to understand the issues that confronted the framers of 1787, the present narrative ends just before the convocation of the Federal Convention.

I shall consider the cooperative or non-cooperative behavior of colonies and states in three arenas: contributions of soldiers and money in wars against the Dutch, Indians, and the British, as well as in the suppression of domestic insurrections; participation in the non-importation, non-exportation, and non-consumption movements directed against Great Britain; and trade relations among the states after independence. I draw heavily on a

This chapter is adapted from a work in progress devoted to a comparison between the Federal Convention of 1787 and the first French Constituent Assembly of 1789–91. I thank the editors for their comments on an earlier draft. Special thanks are due to Jack Rakove for his incisive critical comments on the draft. He should not be held responsible for any mistakes that remain.

J. Elster (✉)
Department of Political Science, Columbia University, New York, NY, USA

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_7

remarkable early study by Arthur Schlesinger Sr. and, on the theoretical side, on recent work by Keith Dougherty.

I begin by citing some discussions by Benjamin Franklin and James Madison that demonstrate their remarkably sophisticated insights into problems of collective action. Although some earlier writers, notably Aristotle (*Politics* 1261 b) and Hume (1978, pp. 520–21), clearly understood the abstract logic of collective action, to my knowledge these two American writers were the first to apply it to concrete social or political situations. While I do not claim there was any direct connection between their analyses and my narrative, I believe they provide valuable pointers to the mindset of people in the colonies and, later, in the states. Madison's brief comments on *shame* as a motivation for cooperation, for instance, echo the more context-specific observations by George Mason.

* * *

Madison's "Vices of the Political System of the United States" will serve as a good starting point. Written shortly before the Federal Convention, this document focuses on three issues: the weakness of the confederation, the lack of cooperation among the states, and the injustice of state laws. Since the weakness of the confederation was largely an effect of the lack of cooperation among the states, I shall consider only the two other issues. With regard to lack of cooperation, I shall also discuss Franklin's "Reasons and motives for the Albany plan of union" (1754). Later, I consider the many plans for union between 1643 and 1781—some of them realized, others not—in the light of these texts.

Any confederation is vulnerable to collective action problems. In the European Union, the "race to the bottom" in corporate taxation offers an example. Such lack of cooperation among members of a union or confederation can have several sources, as Madison noted:

> It is no longer doubted that a unanimous and punctual obedience of 13 independent bodies, to the acts of the federal Government, ought not be calculated on. Even during the war, when external danger supplied in some degree the defect of legal & coercive sanctions, how imperfectly did the States fulfil their obligations to the Union? In time of peace, we see already what is to be expected. How indeed could it be otherwise? In the first place, every general act of the Union must necessarily bear unequally hard on some particular member or members of it. Secondly the partiality of the members to their own interests and rights, a partiality which will be fostered by the Courtiers of popularity, will naturally exaggerate the inequality where it exists, and even

> suspect it where it has no existence. Thirdly a distrust of the voluntary compliance of each other may prevent the compliance of any, although it should be the latent disposition of all.

I shall discuss these three explanations in turn.

I shall generalize the *first explanation* slightly, as "every general act of the Union must necessarily bear unequally hard on *or yield unequal benefits for* some particular member or members of it". Often, there can be several Pareto-improvements (which would benefit all), each of which benefits some agents more than others. In that case, the resentment or envy generated by the adoption of one solution rather than another could in theory be alleviated by side payments to equalize the gains. If, as Madison suggests, some agents might lose from the change, the winners could compensate the losers. (If they cannot afford to, the change is not worth making.) In practice, such schemes tend to be unworkable, because the magnitude of losses and gains will be hard to assess. They are also, for the reason Madison states in his second argument, likely to be controversial.

In the pre-history of the Federal Convention, I have only come across one attempt to create side payments, to equalize losses rather than gains. When the First Continental Congress worked out a non-exportation agreement in 1774, it granted an exception to the rice growers in South Carolina, on the grounds that the non-exportation of rice, unlike that of indigo, would do little harm to the British. "Low-country indigo growers and up-country provision exporters felt slighted and the whole province was bitterly divided along both sectional and interest lines over the partiality shown to the rice planters. In response to the crisis, South Carolina's proponent of the [Continental] association devised an elaborate scheme whereby these smaller producers could swap a portion of their crop for rice at a fixed ratio of value. In this way *the burden of non-exportation would be shared by all*, at the same time preventing the general economic collapse that would have followed a complete embargo on rice, the colony's premier cash crop".[1] The plan was never implemented, perhaps because of complaints on behalf of "the Hemp Grower, the Lumber Cutter, the Corn Planter, the Makers of Pork and Butter etc.", who might also deserve to be compensated.[2]

Madison's *second explanation* for lack of cooperation cites the tendency of the states to exaggerate or even invent such resentment-generating inequalities. Although he does not say why they would do so, one motive could be to justify non-cooperative behavior. A state's refusal to cooperate in a situation of perfect symmetry—in which all states had equal costs and benefits from

cooperation—would burden it with an opprobrium that could be harmful. Although naked interest may be an acceptable motive in the relations among independent states, it is more difficult to defend in a union of states.

Among his explanations of non-cooperation, Madison does *not* include the standard Prisoner's Dilemma.[3] In that model, one assumes that non-cooperation is a dominant strategy for all agents, who are assumed to be moved only by their self-interest. To achieve cooperation, one must rely on negative or positive incentives, imposed by an external authority.[4] In his *third explanation*, he suggests that the situation can be an Assurance Game[5] rather than a Prisoner's Dilemma. In the Assurance Game, the obstacle to cooperation is not self-interest, as it is in the Prisoner's Dilemma, but rather lack of information, or, as Madison says, distrust. If each state were *confident* that the others would comply, it would do so too. Madison's second explanation tends, however, to undercut the third one.

Madison's diagnosis can be usefully supplemented by Franklin's eloquent analysis of the *need* for union, of the *obstacles* to union, and of the best way of *overcoming* the obstacles.

Concerning the *need for union*, Franklin refers to events during King George's War (1744–48), citing the facts that "the assemblies of six (out of seven) colonies applied to, had granted no assistance to Virginia, when lately invaded by the French, though purposely convened, and the importance of the occasion earnestly urged upon them; that one principal encouragement to the French, in invading and insulting the British American dominions, was their knowledge of our disunited state, and of our weakness arising from such want of union; and that from hence different colonies were, at different times, extremely harassed, and put to great expense both of blood and treasure, who would have remained in peace, if the enemy had had cause to fear the drawing on themselves the resentment and power of the whole".

Concerning the *obstacles to union*, Franklin refers to the experience of

> one assembly waiting to see what another will do, being afraid of doing more than its share, or desirous of doing less; or refusing to do any thing, because its country is not at present so much exposed as others, or because another will reap more immediate advantage [...] When it was considered that the colonies were seldom all in equal danger at the same time, or equally near the danger, or equally sensible of it; that some of them had particular interests to manage, with which an union might interfere; and that they were extremely jealous of each other; it was thought impracticable to obtain a joint agreement of all the colonies to an union, in which the expense and burthen of defending any of them should be divided among them all; and if ever acts of assembly in

> all the colonies could be obtained for that purpose, yet as any colony, on the least dissatisfaction, might repeal its own act and thereby withdraw itself from the union, it would not be a stable one, or such as could be depended on: for if only one colony should, on any disgust withdraw itself, others might think it unjust and unequal that they, by continuing in the union, should be at the expense of defending a colony which refused to bear its proportionable part, and would therefore one after another, withdraw, till the whole *crumbled into its original parts* (my italics).

I note, for future reference, that all the obstacles to cooperation that Franklin enumerates are temporary and *conjuncture-dependent*. They are not structural, that is, rooted in the financial or geographical situation of the states. Later I cite a text by Madison that emphasizes these structural and permanent obstacles.

I discuss "crumbling" or unraveling mechanisms later. Here I shall only cite Franklin's proposal for *overcoming the obstacles*: the union must be established and enforced by the British Parliament. Needless to say, this solution was not available when the problem was how to coordinate fight against the British rather than against the Indians and the French.

Both Madison and Franklin considered problems of coordination among the colonies. Madison also considered what he saw as the internal flaws of the colonies—the injustice, the mutability, the multiplicity, and the impotence of their laws. He found the causes of these flaws partly in the representative bodies and partly in the people themselves. Concerning the former, he wrote that "Representative appointments are sought from 3 motives. 1. ambition 2. personal interest. 3. public good. Unhappily the two first are proved by experience to be most prevalent. Hence the candidates who feel them, particularly, the second, are most industrious, and most successful in pursuing their object: and forming often a majority in the legislative Councils, with interested views, contrary to the interest, and views, of their Constituents, join in a perfidious sacrifice of the latter to the former". Concerning the constituents, he distinguished among three motives that might restrain their interests and passions: "1. a prudent regard to their own good as involved in the general and permanent good of the Community. This consideration although of decisive weight in itself, is found by experience to be too often unheeded. It is too often forgotten, by nations as well as by individuals that honesty is the best policy. 2dly. respect for character. However strong this motive may be in individuals, it is considered as very insufficient to restrain them from injustice. In a multitude

its efficacy is diminished in proportion to the number which is to share the praise or the blame"; and "3rdly Religion". On various grounds he also discards the last as an effective counterforce to interest and passions. I shall not discuss religion, but comment on the two other possible counterforces.

The "prudent regard to their own good as involved in the general and permanent good of the Community" is a puzzling phrase. I shall ignore the first words and focus only on the concern for the general and permanent good of the community. This idea corresponds to what the French moralists referred to as *reason*, as distinct from both *interest* and *passion*. Morton White has argued that the American framers adopted an approach to human nature very similar to that of the French moralists.[6] By and large, they did not believe that citizens in general were much swayed by reason, however much they believed themselves to be thus motivated.

By "character" Madison meant "reputation". The explanation of "respect for character" in terms of "praise or blame" shows that he is referring to social norms and thus ultimately to *emotions* of pride and shame. Elsewhere, when he discusses emotions or passions, Madison usually has in mind something like a "violent inclination", a temporary preference reversal caused by strong emotions of anger, hatred, fear, and the like. To prevent impulsive actions that these might induce, one can *pit emotion against emotion*, counting on feelings of pride and shame to neutralize anger or fear, as when soldiers are kept from fleeing the enemy by the shame they would feel before their peers. Madison plausibly asserts that the strength of the counterforce diminishes when the praise or blame is *shared by* (directed toward) many people, but does not mention the fact that it increases when the praise or blame is *expressed* by many people. For this reason, *publicity* can be a crucial element in collective action (see below).

To illustrate the causes and the effects of bad policies, Madison cites laws favoring debtors in the relation to creditors. Concerning the causes, he asks whether it is "to be imagined that an ordinary citizen or even an assembly-man of R. Island in estimating the policy of paper money, ever considered or cared in what light the measure would be viewed in France or Holland; or even in Massts or Connect.? It was a sufficient temptation to both that it was for their interest: it was a sufficient sanction to the latter that it was popular in the State; to the former that it was so in the neighbourhood". Interest speaks louder than what Madison was to call "the mild voice of reason" (*Federalist* # 42), and is also unaffected by what one might call reputation-at-a-distance. By contrast, as we shall see, the concern for reputation can counteract interest in local settings.

Madison discussed both inter-state and intra-state effects of such legislation: "Paper money, instalments of debts, occlusion of Courts, making property [notably tobacco] a legal tender, may likewise be deemed aggressions on the rights of other States. As the Citizens of every State aggregately taken stand more or less in the relation of Creditors or debtors, to the Citizens of every other States, Acts of the debtor State in favor of debtors, affect the Creditor State, in the same manner, as they do its own citizens who are relatively creditors towards other citizens." One might say, perhaps, that such laws generate *injustice* within the state and *inefficiency*, caused by free riding, among the states.

* * *

The history of America from 1643 to 1787 shows a series of attempts, some of them successful, to coordinate the actions of the colonies and later of the states. Earlier efforts sometimes shaped later ones, up to 1787. Before discussing analytical issues, it may be useful to list some of the most important or interesting efforts[7]:

- In 1643, the colonies of Massachusetts, Plymouth (part of today's Massachusetts), Connecticut and New Haven (part of today's Connecticut) formed the United Colonies of New England, mainly to fight against Indians and the Dutch colonies of New Netherland.
- In 1753, Governor Shirley of Massachusetts called for a "union among all the colonies" to coordinate war efforts against Canada. The call was not heard.
- In 1754, 11 colonies met in Albany and adopted a plan of union "for their mutual defence and security and for extending the British Settlements in North America". The plan was turned down on both sides of the Atlantic. According to Franklin, the moving spirit of the Albany Congress, "the [colonial] Assemblies did not adopt it because they all thought there was too much of the [royal] *prerogative* in it; and in England it was judg'd to have too much of the *Democratic*".
- In 1765, the Stamp Act Congress met in New York City, with delegates from nine of the colonies, and adopted a Declaration of Rights and Grievances, submitted as a petition to Parliament, in which they protested against the Stamp Act and insisted on their rights as Englishmen. The Stamp Act was repealed, but as a result of popular resistance rather than because of the Declaration. Yet the Congress may have encouraged the resistance.

- By the fall of 1769, non-importation agreements, to protest against the Townshend Acts, had been adopted in all colonies except New Hampshire. This occurred in a decentralized process, not imposed by a congress of the colonies. The agreements were soon to collapse.
- In 1774, all colonies except Georgia met in Philadelphia to consider how to respond to the Boston Port Act and the Intolerable (or Coercive) Acts. They constituted themselves as the (First) Continental Congress and created an intercolonial association for non-importation, non-exportation, and non-consumption. When Joseph Galloway proposed to create a union between Britain and the colonies, Congress voted six colonies to five, with one divided, to defer further consideration. Whether this vote on a procedural issue reflected substantive preferences is uncertain.[8]
- In 1775, Franklin proposed a draft at Congress for Articles of a Confederation of the colonies. Congress rejected a proposal that a day be set aside to consider the plan.
- In 1776, the 13 colonies declared their independence from Great Britain in Congress, implicitly committing them to contribute money and soldiers to the war.
- In 1776, John Dickinson presented a draft for Articles of a Confederation to Congress.
- In 1777, the Congress adopted a heavily revised version of the Dickinson plan.
- In 1781, Maryland ratified the Articles as the last state.

Drawing on these events, I begin by considering successes and failures of *decentralized* cooperation, mainly in the economic domain. I first provide a narrative summary and then discuss some of the mechanisms that caused success or failure. Next, I discuss the creation of *centralized* bodies and their success or failure in imposing cooperation in the economic, political, and military domains.

Snowballing and Unraveling of Collective Action

Non-importation, non-exportation, and non-consumption movements occurred in three waves, 1765–66, 1767–70, and 1774–75.[9] They involved *nested collective action*, among and within the colonies. Merchants, producers, and consumers in one state would cut their trade, production, and consumption only if they were confident that other colonies did the same.[10]

Also, they might be reluctant to cooperate if the burdens of cooperation seemed to be unequally distributed among the states.[11] At the same time, they would not cooperate unless they were confident that other agents in their own colony did. For this purpose, enforcement mechanisms were essential. In this case, too, they might object to unequal burdens of cooperation across groups within the state, as in the conflict between indigo and rice planers in South Carolina cited earlier.

To understand the dynamics of boycott, we may distinguish among the *snowballing* of cooperation (1765–66, 1767–70), the *unraveling* of cooperation (1770), and the *orchestration* of cooperation (1774–75).[12] The book by Arthur Schlesinger, *The Colonial Merchants and the American Revolution*, published in 1918, provides an outstanding source of insight into these movements, and into the dynamics of collective action more generally. I shall first provide a summary narrative and then consider some of the mechanisms that sustained or undermined cooperation.

In 1764, the American colonies suffered an economic depression, partly because of the collapse of artificial wartime prosperity, but mainly because of the restrictive Revenue Act (or Sugar Act) of 1764. Americans responded by self-imposed sumptuary agreements to cut down on luxury consumption. In eighteenth-century Massachusetts, for instance, social norms required widows and widowers to provide mourners with rings, gloves, and scarves, all of which had to come from England. In 1741, the Massachusetts House of Representatives tried to put an end to these wasteful practices, by forbidding the distribution of scarves and rings and limiting the number of those who could receive gloves to six persons, in addition to the minister and six pallbearers. The legislation was largely ignored.[13] Economic depression achieved what legislation had failed to do. In August 1764, "fifty merchants of Boston set an example [...] by signing an agreement to discard laces and ruffles, to buy no English cloths but at a fixed price, and to forego the elaborate and expensive mourning of the times for the very simplest display". The mourning resolutions were well kept, and estimated savings were £10,000 a year.[14] Other states in New England adopted the same practice. At this stage, these actions were not directed against Britain, but "it did not take the Americans long to perceive that their measures of economic self-preservation might be capitalized to good advantage as political arguments for the repeal of the obnoxious laws",[15] the Stamp Act as well as the Revenue Act. Non-consumption, rather than an end in itself, became a means to non-importation. These actions, a form of passive resistance, were the first steps taken by ordinary Americans to affect British legislation.[16]

As noted, the resolutions of the Stamp Act Congress did not directly bring about the repeal of the Act.[17] Yet "it is possible that the delegates may have agreed that a general boycott through the colonies would be the best method to achieve repeal".[18] After the Congress, which was held in New York City, adjourned on October 25, 1765, the New York merchants called for meetings on October 28 and October 31 at which they "agreed to cancel all outstanding orders on their suppliers in Great Britain, and that they would sell no English goods shipped to them after January 1—until the Stamp Act was repealed.[19] Retailers signed a separate resolution, to protect importers from competition (and perhaps from themselves). These events "triggered a chain reaction".[20] One after another, Pennsylvania, Boston, and other port cities fell into place. These actions occurred more or less in parallel, with a lag of a few months, with acts or threats of physical violence directed against the stamp distributors. After news of violence and boycott reached Britain, English merchants put strong pressure on the government to repeal the Stamp Act. An Act of Repeal passed through both houses of Parliament in March 1766, combined with the Declaratory Act that was intended to save the face of the government while not having any consequences for action. While "most colonists" may have understood the Act in this way,[21] the enactment of the Townshend Acts in 1767 created a new wave of unrest in the colonies.

For my purposes I need not detail the content of these Acts and the reasons why they inflamed the colonists. What matters is that they triggered a new chain reaction of non-importation agreements, in a complicated pattern.[22] First, Massachusetts agreed on non-importation, conditionally on New York and Philadelphia following suit, and then New York, conditionally on Boston and Philadelphia. As Philadelphia refused to follow, New York dropped out. By that time, "non-importation was dead in the North, while the southern colonies remained indifferent. But [...] the conflict between Massachusetts and Britain brought the dead to life. Boston's continuing battle against the customs service led to British measures attacking Massachusetts, which united the colonies in the non-importation movement".[23] Specifically, the southern colonies were aroused by "Parliament's proposal that Americans be brought to England for trial under the treason statute of Henry VIII. That news set going the events which led to the adoption of non-importation agreements by Virginia, Maryland, and South Carolina before the end of the summer of 1769".[24] Around that time, New York and Philadelphia also came around. The boycott was highly successful. "Townshend estimated that [import duties] would bring in

about £ 40,000 a year, but only £ 13,000 was collected in 1768. Thereafter colonial non-importation was effective. The next year collections dropped to a little over £5,000 and in 1770 [. . .] to a little more than £2,500."[25] In New York, the value of imports from England dropped from £482,000 in 1768 to £75,000 in 1769.[26]

By this causal chain, the Townshend Acts brought about the successful non-importation agreements in the colonies. Conversely, "the immediate effect of the repeal of the Townshend revenue act in April 1770 was the collapse of American resistance".[27] The word "immediate" may be too strong. Britain had not, in fact, repealed all parts of the Act, but maintained the tax on tea, mainly, as the king said, because "there must always be one tax to keep up the right". Some colonists wanted full repeal of the Act, whereas others were impatient to resume trading. "The bone and sinew of the non-importation movement were the agreements of the great trading towns of Boston, New York, and Philadelphia. On the action of these towns depended the integrity of the commercial combination. Should the merchants of these towns accept the partial repeal as satisfactory and proceed to revoke their boycott of British importations, this breach in the non-importation dike would render the whole barrier useless."[28] The following weeks saw a process of unraveling that partly matched, in the opposite direction, the snowballing of the previous years. Changing the metaphor, on July 12 New York was the first domino to fall, citing among other things alleged violations of non-importation in Boston. "The patriotic indignation of the other provinces at the defection of New York was splendid to behold. But the merchants throughout the continent realized in their hearts that the prostration of the stalwart pillar of New York would cause the whole great edifice to topple"[29]—which it did by the end of the year.

The last wave of boycotts was triggered by the reaction—in fact, an emotional *overreaction*—of Britain to the destruction of tea in Boston 1773.[30] Britain enacted the four Coercive or Intolerable Acts that closed the port of Boston until the colonists had repaid the costs of the tea, brought the government of Massachusetts under British control, allowed the royal governor to let accused officials stand trial in Britain if he did not think they could get a fair trial in America, and opened for the quartering of soldiers in unoccupied buildings. According to Schlesinger, the effect of the Acts was to polarize society—"merchants at once took their stands with the forces of government and law and order", while for others "the enactment

of the severe punitive acts served [...] to put the greater guilt on the other side".[31]

Schlesinger also cites a thought-provoking comment by Henry Drayton, a "wealthy young South Carolinian who, with fiery zeal, had excoriated [...] the non-importers in 1769":

> The same spirit of indignation which animated me to condemn popular measures in the year 1769, because although avowedly in defence of liberty, they absolutely violated the freedom of society, by demanding men, under pain of being stigmatized, and of sustaining detriment in property, to accede to resolutions, which however well meant, could not [...] but be [...] very grating to a freeman, so, the *same spirit* of indignation [...] actuates me in like manner, *now to assert my freedom against the malignant nature* of the late five Acts of Parliament.[32]

His course was consistent, he asserted: "I opposed succeeding violations of my rights, then, by a temporary democracy, now, by an established monarchy". Fifteen years later, the Comte de Clermont-Tonnerre reversed the temporal order of threats, when he asked his fellow members of the *Constituante*: "You refused to obey armed despotism; are you now going to obey popular effervescence"? According to Tocqueville, such efforts to fight a two-front war are unusual: "it is rare for a man and almost impossible for an assembly to have the ability to alternately make violent efforts in two opposite directions".[33]

Information about the Intolerable Acts triggered parallel and intertwined chain reactions.[34] After the news about the Boston Port Act reached Boston on May 10, 1774, a town meeting on May 13 instructed Samuel Adams to send a circular letter to "all our sister colonies", asking them whether Boston could rely on their "suspending Trade with Great Britain". The circular reached the colonies with various degrees of delay. New York and Virginia received news about the Act before the arrival of the circular letter, and took immediate actions that were later modified in light of the circular. The Carolinas seem to have taken their cue from Virginia. In the end, all the colonies except Georgia agreed to send delegates to an intercolonial congress that would meet in Philadelphia in September to discuss which measures to take in reaction to the Intolerable Acts. The First Continental Congress met on September 5, 1774, in Philadelphia.

Orchestration of Collective Action

Although only five states sent delegates with instructions to adopt non-importation and non-exportation, on October 20 the Congress unanimously created a Continental Association for the organization and enforcement of a boycott. In addition to laying down the rules for non-importation and non-exportation, it imposed severe sanctions on violators. If a trader used the scarcity of goods to increase prices beyond the norm of the previous 12 months, "no person ought, nor will any of us deal with any such person or his or her factor or agent, *at any time thereafter, for any commodity whatever*" (Article 9, my italics). If the committee that was to be chosen in "every county, city, and town" to monitor "all persons touching this association" decides that a person has violated the rules, it shall "forthwith cause the truth of the case to be published in the gazette; to the end, that all such foes to the rights of British-America may be publicly known, and universally contemned as the enemies of American liberty; and thenceforth we respectively *will break off all dealings with him or her*" (Article 11, my italics). Finally, "We do further agree and resolve, that we will have *no trade, commerce, dealings, or intercourse whatsoever*, with any colony or province in North-America, which shall not accede to, or which shall hereafter violate this association" (Article 14; my italics). The document, in fact, "functioned as a constitution for the nation before it was a nation".[35] In the words of T.H. Breen, "local committees throughout America transformed the Articles of Association [...] into a kind of provisional constitution".[36]

In his colony-by-colony account of the work of the association up to the breakout of armed hostilities in April 1775, Arthur Schlesinger found that the rules were very effectively enforced. Although loyalists tried to established counter-associations, their efforts failed, "for the reason that every signer [of the counter-association] at once exposed himself to the wrath of the radicals".[37] The regulations concerning non-consumption were hard to administer, but committees solved the problem by requiring shopkeepers to produce a certificate that the goods they sold had been bought before December 1. It was expected that regulations about simplicity in mourning would be enforced by "friends and neighbours [manifesting] their disapprobation [...] by declining to attend the funeral". "Trials of offenders by the committees of inspection bore every evidence of being fair and impartial hearings, although mistakes were occasionally made". The Southern colonies "without exception resorted to extreme

measures against the merchant-creditors", for instance, by authorizing the committee of inspection to halt proceedings against debtors. A demand even "arose for a boycott against merchants who used excessive caution in extending credit". In Virginia, a man who rented a flat from a person who had condemned the association was obliged to give it up.[38]

Motivations and Mechanisms

At this point, we may pause to reflect on the remarkably passive and non-violent character of the American strategies from 1765 to 1775. By definition, non-importation, non-exportation, and non-consumption are *non-actions*, abstentions from acting.[39] With regard to the treatment of violators, the main reaction stipulated in the Articles of the Association was social *ostracism*, also a non-action, not physical or economic punishment. As noted above, suspected violators were mostly given due process. As shown by the last example in the previous paragraph, ostracism could also extend to non-ostracisers. A person who had worked as a groom for a loyalist family was "drummed and fiddled out of the town, with a strict prohibition of being seen in it again", although some bystanders had asked for tarring and feathering.[40] At this stage that harsh physical punishment, which was frequently imposed during the war, seems to have been less common.[41] "If the insurgents who supported the committees had been as violent as the loyalists claimed, they would have responded by destroying property and endangering the lives of 'many honest worthy persons.' Nothing of the sort occurred."[42] In South Carolina, "some association opponents were menaced [...] but when 'after a little cool reflection' men submitted, it was not so much for fear of violence as of 'a torrent of popular Opinion and perhaps resentment".[43] Ostracism—a show of contempt—can of course be horribly painful for the target, and cause what has been called a form of "civic death".[44] In America, however, the strict rules of the association that violators should be shunned *forever* do not seem to have been respected. Redemption was possible, if the violators showed contrition and promised to respect the rules in the future.[45]

As noted earlier, Madison had little faith in the power of "character" to sustain virtuous behavior, at least in the form of reputation-at-a-distance. In one of the most incisive analyses of the cement needed to sustain non-importation in 1770, George Mason appealed both to character and to interest:

> The Sense of Shame & the Fear of Reproach must be inculcated, & enforced in the strongest Manner; and if that can be done properly, it has a much greater Influence upon the Actions of Mankind than is generally imagined. Nature has impress'd this useful Principle upon every Breast: it is a just observation that if Shame was banished out of the World, she wou'd carry away with her what little Virtue is left in it. The Names of such Persons as purchase or import Goods contrary to the Association should be published, & themselves stigmatized as Enemys to their Country. We shou'd resolve not to associate or keep Company with them in public Places, & they should be loaded with every Mark of Infamy and Reproach. The Interest, too of the Importer may be made subservient to our Purpose; for if the principal People renounce all Connection & Commerce for ever with such Merchants, their Agents & Factors, who shall import Goods contrary to the Tenor of the Association. They will hardly venture to supply their worst Customers with such Articles, at the Hazard of losing their best.[46]

The three boycott movements of 1765–66, 1769–70, and 1774 differed in many respects. What they had in common, to some extent, was that they rode on waves of *conditional cooperation*. Individuals looked to what other individuals did before deciding on what to do, local communities looked to other communities in the same colony, and colonies to other colonies. The processes were complex and defy summary, but I shall try to note some recurring features. In doing so, I rely on two recent books by T. H. Breen.

Obstacles to Cooperation

To the extent that agents had Prisoner-Dilemma preferences, raw self-interest might incline them to non-cooperation. In the words of a contemporary writing in 1774, the previous effort of non-importation had collapsed because "it stood on a rotten and unsolid basis. It was erected wholly on the *virtue* of the merchants, and rested its whole weight solely on this prop".[47] Yet as George Mason pointed out, the *interest* of the merchants in retaining their customers might also induce them to go along. In 1766, "Boston merchants were reluctant to follow [the Philadelphia merchants], but after attacks on them in the popular party newspaper, and threats of non-consumption agreements, they adopted an agreement on 9 December".[48] In some cases, the merchants might also profit from the fact that a boycott would increase the prices on the goods they had stored up, perhaps in the anticipation of non-importation.[49]

As I have noted, many pledges to non-importation and non-exportation were conditional on the agreement of other colonies. It is hard to tell whether colonies that pledged conditional cooperation did so because they had Assurance-Game preferences or because, having Prisoner-Dilemma preferences, they did not want to join and counted on other colonies refusing. In 1768, the Boston merchants cannot have been unaware of the conservative leanings of the merchants in New York and, especially, in Philadelphia. By pledging conditional cooperation they could escape opprobrium and yet count on the conditions not being satisfied.

Considering now other social groups, notably the farmers who made up 90% of the free population, Breen suggests that they had Assurance-Game preferences. The main obstacle to common action that they faced was not self-interest, but the fact "they did not *know* [. . .] whether other Americans shared [their constitutional assumptions] or, if they professed to do so, shared them with the same sincerity".[50] Or again, "Until the colonists forged a greater sense of confidence that other colonists living in other places could be trusted to forgo British imports, they found it hard to translate rhetoric about the renunciation of the market into genuine self-denial and seriously to *join utter strangers* throughout America in resisting a powerful military adversary".[51]

In addition, Britain made it deliberately difficult for the colonists to unite. Celebrating the second anniversary of independence, David Ramsay wrote that "it was the interest of Great Britain to encourage our dissipation and extravagance, for the two-fold purpose of *increasing the sale of her manufactures* and of *perpetuating our subordination*. In vain we sought to check the growth of luxury, by sumptuary laws; every wholesome restraint of this kind was sure to meet with the royal negative".[52] The British wanted to prevent the Americans from escaping the luxury trap by adopting a self-denying ordinance, because if they did so they would hurt their British corrupters. As noted earlier, the royal veto was bypassed by non-consumption agreements. The British might also seek to prevent communication among the colonies, by their control over the postal service. "As long as the British were in a position to interfere with the free flow of communication, they could keep Americans ignorant about the political activities of other Americans."[53]

During the French-Indian wars, lack of unity among the colonies was an obstacle to the military effort. Governor Shirley of Massachusetts urged "the necessity of a union among all the Colonies", imposed by the Crown.[54] "When tensions along the Ohio increased in 1754 [the governor of

Virginia] warned the board [of trade] that an 'Act of Parliament' was necessary to 'compell the Colonies to contribute to the Common Cause, independently of assemblies'."[55] As noted earlier, the Albany Congress of 1754 was in large part motivated by the need to overcome free riding by individual colonies. Yet Britain might have reasons to fear as well as to welcome union among the colonies. "One of the major reasons for British lack of enthusiasm for the Albany Plan of Union was the concern, voiced to [Prime Minister] Newcastle by the Speaker of the House of Commons, that a bill for colonial union would encourage considerable debate over the 'ill consequence to be apprehended from uniting too closely the northern colonies with each other, an Independency upon this country to be feared from such an union'. [...] Those who advocated an all-out military effort against the French sensed that the mother country was constrained by fear lest the colonies become too powerful."[56] Governor Shirley protested, to no avail.

Did Britain foster a deliberate divide-and-rule policy toward the colonies? Generally speaking, failures of collective action can arise not only from internal divisions and distrust among the members of an oppressed group, but also from deliberate attempts by their oppressors to foster distrust by treating them differentially.[57] According to James Nelson, before 1765 "divide and conquer was never a strategy because no one ever thought the colonies would be united enough to need dividing".[58] As early as 1705, Francis Makemie, the founder of Presbyterianism in America, wrote that the British need only "Maintain and propagate the distinct Governors and Governments [...] and Emulation, Division, Heats and Animosities [...] backed by Pride and Envy, will keep them asunder from uniting under a single head, to the prejudice of England".[59] In other words, the natural jealousy and distrust among the colonies that Franklin and Madison were to diagnose in 1754 and 1787 would be sufficient to keep them disunited. Britain had no need to play them out against each other by selectively offering favors to some colonies. Edmund Morgan claims, however, that in 1770, Great Britain, "following a naïve 'divide-and-conquer policy", targeted Boston while "carefully [refraining] from investigating opposition to its authority in other colonies".[60] I have not done the archival research that would be necessary to determine British intentions on this point. As always, one needs to distinguish *divide et impera* (an intention) from *tertius gaudens* (a mere effect).[61] From the fact that the British benefited from conflicts between or among the colonies, one cannot conclude that they deliberately instigated them.

Overcoming the Free-Rider Problem

From obstacles to collective action, I now turn to facilitating conditions. The most important is *publicity*: seeing and being seen, or knowing and being known (Elster 2017).

On the one hand, seeing or knowing that others cooperate can trigger a *quasi-moral norm* of conditional cooperation: if I observe or infer that others do their share, it is only fair that I do mine.[62] (If I observe that they don't, I have no obligation to do so either.) In this respect, newspapers were crucial. In 1769, "it was not unusual [...] to encounter in the newspapers of South Carolina and New York detailed stories recounting how the people of Boston or Pennsylvania had sustained the boycott. [...] As [the governor of New York] observed, 'the chief tendency of them [the newspapers] is to encourage Union among the Provinces.'"[63] A poignant instance—almost a natural experiment—in which people could learn about the commitment of others by observing their behavior occurred when, in early September 1774, a rumor arose that British forces had bombarded and destroyed Boston. While entirely untrue, the rumor spread quickly through New England and caused thousands of people to take arms and march on Boston. "No one had known in advance whether scattered communities from New Hampshire to Connecticut would volunteer to sacrifice – even accepting the possibility of death itself – for a common cause. Now they knew."[64] We may think of these protesters as *first movers*, who by their unconditional behavior trigger the conditional cooperation of others. The first movers may be moved by principle, or perhaps more frequently by emotion. Their behavior and that of their followers hardly supports Gouverneur Morris's claim that "ordinary people had no moral but their interests".[65] In fact, the consensus then and today seems to be that his characterization applied mainly to the merchants.

On the other hand, *knowing that others know* whether you are cooperating, and that they will express disapproval if you are not, can trigger a *social norm* of cooperation, through the fear of naming, blaming, and shaming.[66] The anonymous and impersonal character of life in the large cities made it "hard to shame particular men and women who imported British goods. In these matters, the public needed guidance. Not surprisingly, the popular press provided the remedy".[67] Article 11 of the Continental Association called for the names of the offenders to "be published in the gazette". Examples abound from all three waves of non-importation.[68] In South Carolina in 1769, names of subscribers to a non-consumption

agreement was placed in a register, which anyone could examine. "What few foresaw was that insistence on precision was intended more to punish non-subscribers than to identify the colony's virtuous consumers. The names of the resisters [...] appeared in the newspapers."[69]

Confederations

Unions of colonies and unions of states differ in that the former relate to an external hegemon. In 1643, the hegemon—Britain—was on the brink of civil war, and the four united colonies were left to themselves for nearly a decade. Also, sheer urgency prevented consultation. As one exponent of the United Colonies of New England put it, "If we in America should forbear to unite for offence and defence against a common enemy till we have leave from England our throats might be all cut before the messenger would be half seas through".[70] In 1754, as we have seen, Britain tabled the Albany plan of a union. In this Section I shall consider two actual institutional structures (initiated in 1643 and 1774) and two plans proposed by Benjamin Franklin that never left the planning stage. The purpose is to bring out some dilemmas of confederations that figured prominently in the thinking of the 1787 framers. The main issues are the *representation* of the states in Congress, the *voting* procedures in Congress, and the *contributions* by the states to the central treasury.

The New England Confederation of 1643

This union included one large colony, Massachusetts, and three substantially smaller ones, Connecticut, Plymouth, and New Haven. As stated in the Articles of Confederation, it was organized for both defensive and offensive purposes, with the proviso that any war had to be "just".[71] The governing body was made up of commissioners, who met once a year and from time to time as needed. Each colony sent two commissioners, and decisions required a supermajority of six.[72] If six commissioners could not agree, the matter would be referred to the General Courts (legislatures) of the colonies. In times of danger, Massachusetts would provide 100 soldiers and each of the others 45 or less, as required by proportionality. According to a literal reading of the Articles of Confederation, the commissioners, if six of them agreed, had full powers to decide and execute military action. During the first decade of the confederation, "it looked as if the lesser members sought to promote the conception that a super-government

transcending the General Courts was the intention of the founders",[73] whereas Massachusetts was reluctant to surrender its sovereign powers. "The violent controversy [in 1653–55, concerning an offensive war against the Dutch], with New Haven, representing the small colonies, pitted against Massachusetts, raised for the first time in American history the question whether the central authority or the individual colonies were supreme."[74] On that occasion, "Massachusetts, the most remote from the danger of war and upon whom the burden would be the greatest, stubbornly refused" to join the smaller colonies in an offensive war against the Dutch.[75]

In the Revolutionary War and in the suppression of Shays' rebellion, one also observed the reluctance on the part of some colonies or states against fighting when the danger was not imminent or close.[76] According to Keith Dougherty, this problem is inherent in any confederation without a strong central authority. Unless a member state derives some private benefits from cooperation, it will stay on the sidelines.[77] Here I shall only comment on the unequal size of the member colonies, an issue that cast a shadow down to 1787. A large unit might demand a larger number of delegates to the council of the confederation or, failing that, a veto. During a conflict in 1649, the Massachusetts commissioners argued that by virtue of the fact that the colony bore "almost five to one in the proportion of the charge with any one of the rest", it was entitled to three commissioners.[78] In the revived confederation of 1672 (including only Massachusetts, Connecticut, and Plymouth), a decision required the agreement of five out of six commissioners.[79]

If, as in the present case, a confederation has only *one* large member and several small ones, the latter may with some justification fear the dominance of the former, even if the formal rules place them on an equal footing.[80] If there are *several* large members, their tendency to dominate will depend on their commonality of interest. They obviously share an interest in representation in Congress being proportional, but they might have entirely different economic interests. Even when those interests are too different to generate a "coalition of the large", the smaller members may argue for equality of representation, alleging fear that the larger members will oppress them. The fear may be groundless, and the allegations hypocritical, yet the argument can be politically efficacious.

Franklin's Two Plans

Benjamin Franklin variously proposed to base representation in Congress on population or on contributions to the common funds of the confederation. In the Albany plan that he penned,[81] the colonies were initially (in the first 3 years) to be represented in the Congress of the confederation in the following proportions:

Massachusetts	7
New Hampshire	2
Connecticut	5
Rhode Island	2
New York	4
New Jersey	3
Pennsylvania	6
Maryland	4
Virginia	7
North Carolina	4
South Carolina	4
Total	48

Although Franklin did not explain how he arrived at these numbers, they seem to be based on population, including slaves. In 1750, the white population of Rhode Island was larger than that of South Carolina, 30,000 versus 24,000, whereas the totals including slaves were 33,000 and 64,000.[82] The ratio of the largest to the smallest number of delegates was 3.5 to 1, whereas the corresponding population ratio (including slaves) was about 8.5 to 1 (Virginia with 231,000 to New Hampshire with 27,000 inhabitants). This tendency to underrepresentation of large members is very common in federal systems (or in their upper houses). However, in the plan Franklin presented to the Second Continental Congress in 1776, he proposed strict proportionality between representation and population (presumably also including slaves).

In the Albany plan, Congress jointly with a President appointed by the crown was to be granted the power to levy taxes and duties, and to requisition payments from the treasuries of the colonies. In the 1776 plan Franklin proposed that "All Charges of Wars, and all other general Expences to be incurr'd for the common Welfare, shall be defray'd out of a common Treasury, which is to be supply'd by each Colony in proportion to its Number of Male Polls between 16 and 60 Years of Age; the Taxes for

paying that proportion are to be laid and levied by the Laws of each Colony". Contributions and representation should both be strictly *proportional to population*. By contrast, in the Albany plan Franklin had proposed that representation should become *proportional to contribution*, once "the proportion of money arising out of each Colony to the General Treasury can be known", subject to an upper limit of seven and a lower limit of two delegates per colony. There are no such limits in the 1775 plan.

The Albany plan tacitly presupposes that voting in Congress would be by simple majority. Franklin was more explicit concerning the quorum, which was to consist "of twenty five members, among whom there shall be one or more from a majority of the Colonies". In 1775, he retained the requirement that a majority of delegates be present, but omitted the requirement that a majority of colonies be represented. Together with the absence of upper and lower limits on the number of delegates, this omission suggests that Franklin in 1775 was less concerned with the rights of the colonies or incipient states than he had been in 1754.

An American Mystery

Although never implemented, Franklin's plans are interesting in the way they identify and address some of the major issues in organizing a union of colonies or states. Before I consider how these issues were debated and ultimately resolved by the Second Continental Congress, I need to explain, as best I can, the adoption in September 1774 of the ground rules for both the First and the Second Congress. I say, "as best I can", because the process is shrouded in a mystery that is hard to penetrate. With one possible exception, the historians I have consulted have not addressed the issue. The question is simple: how should the delegates *decide how to decide*? The appointment of the 12 delegations was a pretty haphazard process. It seems that each colony sent as many delegates as it could easily afford. Not only were the delegations of unequal size, but the numbers were not in any way proportional to the population or wealth of the respective colonies. Once the delegates were in place, they had to decide on how they were to decide in the future. The options were:

- Each colony will cast one vote.
- Each delegate will cast a vote.
- Each colony will cast a number of votes that is proportional to its population, wealth, or some combination of the two.

There were obvious objections to each proposal, clearly stated by John Adams in his Diary:

> [i] If We vote by Colonies, this Method will be liable to great Inequality and Injustice, for 5 small Colonies, with 100,000 People in each may outvote 4 large ones, each of which has 500,000 Inhabitants. [ii] If We vote by the Poll, some Colonies have more than their Proportion of Members, and others have less. [iii] If We vote by Interests, it will be attended with insuperable Difficulties, to ascertain the true Importance of each Colony. – Is the Weight of a Colony to be ascertained by [iiia] the Number of Inhabitants merely – or [iiib] by the Amount of their Trade, the Quantity of their Exports and Imports, or [iiic] by any compound Ratio of both. This will lead us into such a Field of Controversy as will greatly perplex us. Besides I question whether it is possible to ascertain, at this Time, the Numbers of our People or the Value of our Trade. It will not do in such a Case, to take each other's Words. It ought to be ascertained by authentic Evidence, from Records.[83]

Adams does not mention whether, in cases (iiia) or (iiic), slaves would be counted, fully or partially. The question would come up on the Second Continental Congress, with respect to the contributions of the colonies to the central treasury and in 1787, with respect to their representation in Congress.

On September 6, the Congress "Resolved, That in determining questions in this Congress, each Colony or Province shall have one Vote. The Congress not being possess'd of, or at present able to procure proper materials for ascertaining the importance of each Colony".[84] Virtually all commentators content themselves with affirming that Congress "resolved" or "agreed" to adopt the principle of "one colony, one vote" for the proceedings of the Congress, without specifying *how* the decision was made.[85] One tantalizing exception is a 1942 biography of John Rutledge from South Carolina by Richard Barry, a "semi-scholarly book" according to a recent historian.[86] In Barry's story, Rutledge was outwitted by Samuel Adams on the first days of the Congress:

> With the initial business went a motion that the votes should be counted by colonies, not by individuals. Virginia had seven votes [...], South Carolina five, and Massachusetts only four. There were more northern colonies than southern, but if Pennsylvania were counted with the South, and she leaned that way, the southern interest would command more individual votes. Also, individually there were more moderates than radicals. It was to the advantage

> of the South as well as to that of the moderates to vote by delegates. Apparently only Sam Adams saw this from the start. He consorted day and night before the opening with his adoring pupil, Christopher Gadsden [the "Sam Adams of South Carolina"]. Then in the first debate, Gadsden [...] was on his feet with a blunt call for a vote by colonies. Sam Adams was rushing John Rutledge off his feet at the first impact by the simple expedient of leading off with Rutledge's associate. [...] The conservatives in Congress were whipped by that first bold move which came before the issue was defined. The decision to vote by colonies was carried by the narrow margin of two votes.[87]

At this time, there were 11 colonies present and 50 delegates. A margin of two individual votes implies 26 to 24. A margin of two colonies, assuming that one of the even-numbered delegations of Rhode Island and Pennsylvania was prevented by a tie from casting a vote, implies 6 to 4. I tend to believe that only the first margin counts as being "narrow", but that is obviously a matter of judgment. If I am right, Adams manipulated the vote by individuals to make congress adopt the vote by colonies.

Independently of how it was made, the decision to adopt "one colony, one vote" had momentous consequences, certainly in the long run (see chapter "A Political Theory of Constitutional Democracy: On Legitimacy of Constitutional Courts in Stable Liberal Democracies") and arguably in the short run. If Barry is right, a vote by individuals might have led to a reconciliation with Britain, or at least to an attempt:

> The value of the victory to Adams appeared in the first major business, a resolution prepared by Joseph Galloway, of Pennsylvania, providing for an administrative separation of England and America, the colonies remaining under the Crown, yet having full authority to levy taxes, while all equities of the carrying trade would be administered by a joint commission. It was an enlightened proposal, and, if it had been adopted in Philadelphia and ratified in London, would have prevented the war. Rutledge favored the Galloway plan and was its floor leader, but the radicals under Sam Adams marshalled their forces, denounced it as a Royalist plot, and defeated it by seven colonies against six. Except for Christopher Gadsden, every southern vote, including that of George Washington, as well as a comfortable minority in the North, was for the Galloway plan. If the votes had been by individual delegates, the plan would have succeeded.[88]

Barry does not state explicitly that Adams knew about and took strategic precautions against the Galloway plan. As a universally acknowledged master strategist, Adams was certainly aware of the balance of opinion among individual delegates and may well have expected some proposal of this kind. If my interpretation is correct, on September 6 he persuaded less sophisticated and informed delegates to vote for a voting system that would not further their preferences.[89]

Representation and Voting in the Continental Congress

The voting rules adopted on September 6, 1774, remained in force when, after Independence, the Continental Congress debated and voted on the voting rules to be adopted in the Articles of Confederation. Although Franklin's proportional voting scheme was not debated, he took an active part in the discussions. Responding to a proposal that all states should have an equal vote in matters of "life and liberty", but that in money matters votes should be proportional to population, he said (according to Jefferson's notes) that "the votes should be so proportioned in all cases. He took notice that the Delaware counties had bound up their Delegates to disagree to this article. He thought it a very extraordinary language to be held by any state, that they would not confederate with us unless we would let them dispose of our money. Certainly if we vote equally we ought to pay equally: but the smaller states will hardly purchase the privilege at this price".[90]

On June 12, 1776, a committee, of which John Dickinson was the dominant member, proposed a first draft of Articles of Confederation, which gave wide powers to Congress. "The sole restraint upon the power of Congress was that it might not lay taxes and duties, which was logical enough if the American Revolution was in any sense a revolt against taxation by an external and superior political agency".[91] Whether logical or not—the analogy between Great Britain and the Congress is pretty halting—the restraint was certainly psychologically intelligible and turned out to have momentous consequences. In the subsequent debates, the Dickinson draft was somewhat diluted, the preponderance of power being retained by the states.[92]

The draft also included the clause, "In determining questions each State shall have one vote". The debate on this issue went along predictable lines. "The members of Congress from the larger states developed many

ingenious theories to support their demand for a preponderant influence in the union. [...] Obviously, their arguments were dictated by their desire to give their states a dominant voice in the affairs of the union. Their opponents from the small states knew this and dwelt upon it persistently, for they feared that they would be 'swallowed' by their great neighbors."[93] In the last statement, we should probably replace "feared" by "claimed to fear". The small states, no less than the large ones, simply wanted to have as much influence as possible. The fear was spurious, as James Wilson pointed out: "I defy the wit of man to invent a possible case or to suggest any one thing on earth which shall be for the interests of Virginia, Pennsylvania & Massachusets, and which will not also be for the interest of the other states".[94] Benjamin Rush concurred: "The larger colonies are so providentially divided in situation as to render every fear of their combining visionary. Their interests are different, & their circumstances dissimilar. It is more probable they will become rivals & leave it in the power of the smaller states to give preponderance to any scale they please."[95] As noted earlier, a confederation with *many* large members may be less threatening to the small ones than one with a single large member, as was the case in the Colonial Union of 1643.

The "one state, one vote" clause was retained in the revised draft that Congress adopted on August 20, 1776. Virginia, in one case joined by Pennsylvania, proposed four amendments to take account of population size, but was consistently defeated.[96] In the Articles of Confederation that were finally ratified in 1781, after Congress resolved the question of sovereignty over the Western lands, this clause was also retained and remained in force until 1787. While the clause required the states to vote as delegations, equality of the states would also have been compatible with each state sending the same number of delegates to vote as individuals.[97] Instead, with voting by delegations the states sent varying number of delegates, constrained to be between two and seven.

From representation in Congress, I now turn briefly to its voting procedures.[98] In the 1781 version of the Articles, the quorum was set to nine states. Any changes in the Articles themselves had to be ratified by all states. On an enumerated set of issues, the assent of nine states was required. Otherwise, majority voting was sufficient. It might seem, therefore, as if a simple majority of five to four could be decisive. Congress decided, however, to interpret majority in an absolute sense, requiring the vote of seven states. This procedure was made even more stringent by the fact that if a delegation was tied in its internal vote, it did not contribute to the quorum.

In addition, physical absenteeism raised the bar for decisions even higher. Even though Congress lowered the quorum to seven in 1783, its proceedings were often paralyzed, as described by Thomas Jefferson in a letter to George Washington from May 15, 1784:

> I suppose the crippled state of Congress is not new to you. We have only 9 states present, 8 of whom are represented by two members each, and of course, on all great questions not only an unanimity of States but of members is necessary. An unanimity which never can be obtained on a matter of any importance. The consequence is that we are wasting our time & labour in vain efforts to do business. – Nothing less than the presence of 13 States, represented by an odd number of delegates will enable us to get forward a single capital point.

Contributions

Congress also discussed the bases for contributions to the Union. Whereas representation is zero-sum, contribution is not. All member states of a confederation benefit from internal law and order, and from defense against other nations, but each state would prefer others to bear the main burden. To overcome the free-rider problem, one might link the representation of the states to their contributions just as, on the individual level, one might link the right to vote, or even (as Turgot proposed in France) the number of votes, to the payment of taxes. Thus, on August 1, 1776, Middleton of South Carolina "moved that the Move should be according to what they pay" in taxes.[99] An alternative solution, not based on incentives, would be to link both representation and contributions to population. As we saw, Franklin at various times advocated both solutions. However, with representation decided by the principle "one vote, one state", contribution became a freestanding and independent issue.

The free-rider issue is not the only obstacle to taxation, however, perhaps not even the most important one. Although the confederation obviously needed *some* "general revenue", the states differed regarding their (permanent) interests in what the revenue would fund and, consequently, in the amount of revenue to be raised. In two remarkable memoranda on the interests of each of the 13 states dated February 26 and March 6, 1783, Madison lists a number of reasons why the states might favor or oppose taxation at the level of the confederation. Some states would support it because an impost duty levied by the confederation would spare them predatory imposts levied by neighboring states. The latter states would

oppose it for the same reason. Many states would support "abatements", that is, compensation for their disproportionate losses or expenses during the revolutionary war; other states, which could not make a claim for such losses, would oppose measures to satisfy it. Some but not all states would support revenue to absorb debts incurred during the war. Since the revenue was expected to strengthen the authority of the central government over the Western lands, states that claimed a prior entitlement to these and states that wanted to acquire them had opposing interests. To convey the flavor of Madison's analysis, I shall cite two of his diagnoses:

> Rhode Island as a weak State is interested in a general revenue as tending to support the confederacy and prevent future contentions, but against it as tending to deprive Her of the advantage afforded by her situation of taxing the commerce of the contiguous States. As tending to discharge with certainty the public debts, her proportion of loans interest her rather against it. Having been the seat of the war for a considerable time, she might not perhaps be opposed to abatements on that account. The exertions for her defence having been *previously* sanctioned, it is presumed in most instances, she would be opposed to making a common mass of expenses. In the acquisition of vacant territory she is deeply and anxiously interested.

> Virga. in common with the Southern States as likely to enjoy an opulent and defenceless trade is interested in a general revenue, as tending to secure her the protection of the Confederacy agst. the maritime superiority of the E. States; but agst it as tending to discharge loan office debts and to deprive her of the occasion of taxing the commerce of N. Carolina. She is interested in abatements, and essentially so in a common mass, not only her excentric expenditures being enormous; but many of her necessary ones havg. rcd. no previous or subsequent sanction. Her cession of territory would be considered as a sacrifice.[100]

Given these opposed interests, and the practical obstacles to side payments, it is not surprising that the general revenue plan that Hamilton and Madison prepared in 1783 never received the unanimous approval from the states it needed for adoption.

Compliance of the States with the Requests from Congress

Although Congress used its power to request the states to contribute to the common treasury in proportion to their population, they did not always comply. I shall discuss the issue of non-compliance in a more general perspective, which includes military as well as financial contributions. In doing so, I must proceed differently from what I did in the previous section, since relations among colonies or among states differ from those among individuals. I have argued that to understand how individual Americans were able, on several occasions, to overcome their free-rider problems, we must invoke emotions, quasi-moral norms, social norms, and self-interest (fear of losing customers). When the colonies or states faced a free-rider problem, as they did both during and after the war, their perceived interests were dominant in shaping their behavior.[101] Needless to say, one cannot talk coherently about *the* interest of a state. In a given state, different groups—farmers and merchants, consumers and producers—may differ in their interests. The way in which these interests are aggregated into state *decisions* is often opaque and beyond analysis, although in some cases one can follow the causal chain.

In 1774 and in 1776, the colonies and states benefited from the urgency of the situation. Had they proceeded more leisurely, they might not have reached agreement on the voting rule in the First Continental Congress, nor on the method for calculating the contributions of the states to the common treasury. Writing to Richard Caswell on November 4, 1777, Thomas Burke said that "I deem a time of peace and tranquility, the proper time for agitating so important a concern, but some and not a few, are of opinion that advantage should be taken of the present circumstances of the States which are supposed favorable for pressing them, to a very close connexion". If he alluded to the benefits the small states derived from the voting rule and the Southern states from the contribution rule, he was right. If he thought more impartial solutions would have been found if the states had been closer to "the ideal speech situation", he was almost certainly wrong. In the wake of the Intolerable Acts and the de facto declaration of war with Britain, urgency forged consensus among the delegates.

The actual conduct of war and the finances of the confederation required, however, *compliance* with these and other decisions. If states are motivated mainly by their self-interest, compliance may be difficult to achieve. In considering war, I shall cite examples from the wars against the Dutch, against the French and Indians, against the British and against

domestic insurrections. In 1653, Massachusetts refused to take part in a joint action against the Dutch, both because they had little reason to fear them and because they would have to assume most of the burden. One of Franklin's arguments in 1754 for an intercolonial union was that state legislatures are "at present backward [reluctant] to build forts at their own expense, which they say will be equally useful to their neighboring colonies". Regarding the war against the British, Forrest McDonald writes that "patriotism and the proximity of the enemy proved to bear an almost one-to-one relationship. Men loved their country – or were interested that its 'government' do anything – whenever British troops were in sight, and with rare exceptions only then".[102] He echoes the argument in which Joseph Galloway presented his plan to the Continental Congress: "You all know there were Colonies which at some times granted liberal aids, and at others nothing; other Colonies gave nothing during the war; none gave equitably in proportion to their wealth, and all that did give were actuated by partial and self-interested motives, and gave only in proportion to the approach or remoteness of the danger".[103]

In *warfare*, the common interests of the colonies or states were relatively easy to perceive. Nevertheless, as Franklin pointed out, an individual state might, out of shortsightedness, fear an unfair sharing of the burden, or out of suspicion of other states, refuse to join the common effort. These are changing and conjuncture-dependent sources of interest heterogeneity. In *financial questions*, as Madison pointed out in his 1783 memoranda, the states may not even have (or believe they have) a common interest. Some states would welcome and others would oppose the raising of revenue that would strengthen the central government. In his pioneering study of collective action under the confederation, Keith Dougherty overlooks this more permanent and structural heterogeneity of interests. Once we take account of this fact, Dougherty's basic puzzle appears even more striking: "Since requisitions were voluntary, their failure should be of little surprise to modern scholars. The real mystery is why the states paid any money to Congress" at all.[104]

To resolve the puzzle, Dougherty appeals to the "theory of joint products" according to which a contribution to a public good, such as fight against a common enemy, may also provide private and motivating benefits to some contributors. For my purposes, the most interesting and convincing application of the theory is his analysis of Shays' rebellion. Although the Continental Congress called upon the states to provide soldiers and money to crush the rebellion, most states did not comply. The rebellion was defeated only because the governor of Massachusetts raised an army from

private sources. I am less convinced by Dougherty's analysis of state contributions to the central treasury, because he underestimates or even ignores what I called the structural heterogeneity of state interests that Madison commented on in great detail.

* * *

In conclusion, it is striking that farmers, planters, and artisans were willing to forsake their self-interest, under the influence of quasi-moral or social norms, whereas merchants and politicians were not. Yet even when citizens managed to rally around the common cause, their cooperation was always fragile. Ultimately, it was rooted in the emotions of anger and enthusiasm of the first movers, who triggered the conditional cooperation of other citizens. Yet since emotions notoriously have a short half-life, the cooperation could easily unravel. Nevertheless, while it lasted it could inspire actions that caused British overreactions, which might keep flames burning that might otherwise have subsided by themselves.

The proceedings at the Federal Convention in 1787 were shaped by collective action issues that had arisen over previous decades, notably by the free-riding temptation that was a built-in feature of the non-importation, non-exportation, and non-consumption movements, as well as of the Articles of Confederation, and that the framers successfully managed to eliminate. How did they do so? In my forthcoming work on the Convention, I shall argue that the seaboard elites were galvanized into cooperating by Shays's Rebellion, as the colonies had been galvanized by first the economic and then the military struggle against Great Britain (for a sketch of this argument, see Elster 2012). The inefficiency of the confederation was notorious, but it took a crisis to make it causally efficacious.

Notes

1. Starr (2000), p. 227; my italics.
2. Schlesinger (1968), p. 468.
3. Some of his examples of "Trespasses of the States on the rights of each other" may illustrate this model, an example being "the law of Virginia restricting foreign vessels to certain ports – of Maryland in favor of vessels belonging to her own citizens – of N. York in favor of the same". See also his memoranda from 1783 discussed and quoted below.
4. Under some conditions, agents might be motivated to cooperate out of *long-term* self-interest. With many agents, this can happen if they all adopt the "grim-trigger" strategy: cooperate until one agent defects and never

cooperate thereafter. Whereas this mechanism might conceivably explain the stability of cartels, it is highly implausible as an account of cooperation among states.

5. Dougherty (2003).
6. White (1987), Chap 7; also Elster (forthcoming).
7. The list of proposed plans for union is selective. For a fuller compilation, see Stone (1889), vol. II.
8. According to the *Journal of the Continental Congress*, "[a]ll the men of property, and most of the ablest speakers, supported the motion, while the republican party strenuously opposed it". See also comments by Richard Barry cited later. Ammerman (1974, pp. 58–60) dismisses the Galloway plan as unimportant. The fact that Congress later struck it from its Journal seems to count against this view. The contemporary evidence is summarized in Smith (1976, pp. 112–17).
9. In 1774, some colonists also demanded the non-payment of the crippling debts to British merchants (Schlesinger 1968, pp. 404–5, 414, 416). This proposal was not made part of the platform of the Continental Association.
10. Breen (2004, p. 200). In 1769–70, distrust of the Boston merchants was an important obstacle to non-importation in other colonies (Jensen 1968, pp. 366–71).
11. See Schlesinger (1968), pp. 421–22, and Gould (1986), p. 35, on John Rutledge's qualms in this respect at the Continental Congress.
12. See Elster (2015), pp. 388–97, for analyses and examples.
13. Breen (2004), p. 213.
14. Schlesinger (1968), p. 63.
15. Ibid., p. 77. Once the Stamp Act was repealed, "people again bowed to the custom of expensive funerals" (ibid., p. 86). After the Townshend Acts, frugality came back in the colonies (ibid., pp. 107, 110, 143, 146).
16. Some opponents to non-consumption claimed that concerted actions to punish innocent third parties, the British producers, were illegal. Writing in the *South Carolina Gazette*, "William Wragg [...] argued that it did not follow that a number of persons associating together had a right to do what one man might do, and he said that Parliament had acted on this doctrine in punishing tailors for combinations to increase wages" (ibid., p. 205 n.). (The Americans did not, of course, want to benefit at the expense of the producers in the way the tailors wanted to benefit at the expense of their employers.) See also Maier (1991), pp. 133–34.
17. Weslager (1976), p. 248; Jensen (1968), p. 124.
18. Weslager (1976), p. 242.
19. Ibid., pp. 240–41.
20. Ibid., p. 240.
21. Maier (1991), p. 145.

22. In the following I rely on Jensen (1968) and on the much more detailed narrative in Schlesinger (1968). The main point on which they diverge concerns the causes of the adherence to non-importation in the Southern colonies and its impact on the Northern ones. Not being a specialist, I assume that Jensen's more recent account is the more reliable.
23. Jensen (1968), pp. 272–73. As he explains (ibid., p. 279), "the stationing of the custom boards [for all the colonies] in Boston was one of the major political blunders of the age; in any other American city it would have had less trouble".
24. Ibid., p. 302. Schlesinger (1968) does not refer to this proposal to explain second wave of non-importation. He does, though, cite it when discussing the later wave of non-exportation (*op. cit.*, pp. 397, 416).
25. Jensen (1968), p. 331.
26. Ibid., p. 359.
27. Jensen (1968), p. 329.
28. Schlesinger (1968), p. 217.
29. Ibid., p. 229.
30. On the emotional character of the British reactions, see Ramsay (1990), vol. 1, pp. 88–89; on the emotions they triggered in America, see ibid., pp. 89–90.
31. Schlesinger (1968), p. 309. The fifth Act was the Quebec Act, which "aroused a wider variety of complaints than any of the other four statutes": it established a government without a representative assembly, permitted the continuation of a legal system not allowing in all cases for jury trial, created a feudal system of land tenure that frustrated the hopes of speculators, and by providing for the Catholic religion raised the specter of an established church (Ammerman 1974, p. 11).
32. Ibid., p. 310.
33. Tocqueville (2004), p. 610. Tocqueville was proud of the fact that his great-grandfather Malesherbes was among those who did resist in both directions.
34. Although Schlesinger (1968) and Jensen (1968) provide a great deal of detail, the pictures they draw are still incomplete.
35. Breen (2010), p. 189. Earlier, similar proto-constitutional organizations had arisen at the colony level. By 1770, "the various [non-importation] associations came to serve as social compacts, analogous to the formal constitutions that would be set up by the various colonies in the mid-1770s" (Maier 1991, p. 136).
36. Ibid., p. 18.
37. Schlesinger (1968), p. 477. He adds (ibid., p. 478) that "the failure of the loyalist association was due to the superior organization of the radicals rather than to lack of support for it". The empirical issue seems hard to

resolve: how many sanctioned violators of the regulations out of patriotism or out of fear of being sanctioned for non-sanctioning?

38. The examples in this paragraph are from Schlesinger (1968), pp. 477, 481, 483, 488, 504, 505, 511, 514.
39. True, the coordination of non-actions may require action, which may be why governors demanded that non-importation and non-exportations could be banned as illegal. In many cases, however, the coordination was the spontaneous result of observing what other people were doing.
40. Breen (2010), p. 196; also Breen (2004), p. 26.
41. Irvin (2003) examines about 70 cases of "tar and feather" between 1768 and 1776. In the period before the outbreak of hostilities, the main targets were custom officials, importers, and informers; see also Maier (1991), pp. 128–29. Irvin does not include a single case in which *consumption* was punished in this manner.
42. Breen (2010), p. 211.
43. Maier (1991), p. 122.
44. See Elster (1999), pp. 146–47, 234, for examples.
45. Breen (2004), p. 263; Breen (2010), p. 171; Schlesinger (1968), pp. 495, 556–7, 565, 581.
46. Mason (1970), vol. I, pp. 116–17, see also Franklin (1959–), vol. XVII, p. 202. Mason seems to have thought that importers were shameless, but that customers might be deterred by shame before their peers. On shame as the "false coin" that must substitute for the true coin of virtue, see also Montaigne (1991), p. 715.
47. Cited after Breen (2004), p. 299; my italics.
48. Jensen (1968), p. 129. If George Mason was right, the merchants would be worried more by the loss of income through non-consumption than by the loss of reputation through attacks in the newspapers.
49. Schlesinger (1968), p. 114.
50. Breen (2004), p. 23; my italics.
51. Ibid., p. 200; my italics; see also Breen (2010), p. 103, on this problem of *trust among strangers.*
52. Ramsay (1778), p. 64; his italics.
53. Breen (2010), p. 108. Coastal consumer trade, however, "carried messages from other colonies" (Breen 2004, p. 127).
54. Shannon (2000), p. 71.
55. Ibid.
56. Bumsted (1974), p. 550.
57. See, for instance, Acemoglu et al. (2004).
58. Nelson (2011), p. 64.
59. Cited after Miller (1985), p. 339.

60. Morgan (2012), pp. 47–8. Jack Rakove (personal communication) agrees with Morgan, but adds that "the great puzzle here is why the British doubled down on [the divide and rule] policy in 1775 when it had so obviously failed in 1774".
61. Simmel (1908).
62. On quasi-moral norms, see Elster (2015), Chap. 5.
63. Breen (2004), p. 250. As we shall see shortly, newspapers could also make people realize that they were *being seen*.
64. Breen (2010), pp. 150–51.
65. Wood (1991), p. 27.
66. Elster (2015), Chap. 21.
67. Breen (2004), p. 261.
68. Ibid., pp. 254–55; Maier (1991), pp. 121–22; Schlesinger (1968), pp. 130, 150, 158, 164, 185, 217, 477; Jensen (1968), p. 193.
69. Breen (2004), p. 271.
70. Cited after Ward (1961), p. 37. The same obstacles of time and space prevented American representation in the British parliament.
71. For the role of "just wars" in the Union, see Muehlbauer (2008).
72. Articles of Confederation of 1643, Appendix A to Ward (1961). The voting rule was adopted unanimously by a first meeting of the commissioners and ratified by the legislatures of all the colonies. At that time, unlike later founding moments, "deciding how to decide" was not a contentious issue.
73. Ward (1961), p. 54.
74. Ibid., p. 178.
75. Ibid. Massachusetts claimed, speciously, that the war was not just. "In its puritanical way, [it] had sought to assume moral responsibility to prevent war with the Dutch and their Indian allies. Actually, practical considerations were foremost: the brunt for carrying on a war would rest with MA, the largest member of the Confederation, and the Bay colony would have the least to gain" (ibid., pp. 191–2). According to Muehlbauer (2008), p. 329, "The disparate assessments of danger among the Puritan colonies practically destroyed the Confederation in 1653".
76. As Franklin and Governor Shirley noted, the free-rider problem also arose in the French-Indian wars.
77. Dougherty (2001), pp. 13–14 and *passim*.
78. Hazard (1794), vol. II, p. 199. They added slyly that they would agree to any other colony also having three commissioners if it was willing to bear the same charge as Massachusetts.
79. Articles of Confederation of 1672, Appendix B to Ward (1961).
80. Ward (1961), p. 43, writes that the Hanseatic League, "like the later confederations of the Dutch and the [New England] Puritans [...] was to

suffer from the aggrandizement of the leading province over the lesser states".

81. Reprinted as an Appendix to Shannon (2000).
82. Numbers from https://web.viu.ca/davies/h320/population.colonies.htm
83. Adams (1976a), p. 10; see also Kromkowski (2002), pp. 153–55.
84. They added that "As this [resolution] was objected to as unequal, an entry was made on the journals to prevent it being drawn into a precedent". As we shall see in Chap. 8, this entry had no effect.
85. "The impossibility of fixing the comparative weight of each province [...] induced congress to *resolve*, that each should have one equal vote" (Ramsay 1990, vol. 1, p. 106); the members "*agreed* that the delegates of each province should cast one vote collectively" (Schlesinger 1918, p. 412); "In the end each colony *was given* one vote" (Jensen 1950, p. 59); "it *was agreed* that each colony should have one vote" Burnett (1964, p. 38); "Congress finally *agreed* that each colony should have one vote" (Jensen 1968, p. 492); "the delegates then *resolved* 'that . . . each colony or province should have one vote'" (Jillson and Wilson 1994, pp. 52–53). The authors of the phrases I have italicized do not indicate the voting rule.
86. Hutson (1987), p. 416. Barry's book has no footnotes or other scholarly apparatus. Other scholars are even more dismissive; see for instance Haw (1997), p. vii. However, Forrest McDonald (1979, p. 266), whose judgment I tend to trust, says that Barry's book "has a great deal of data and penetrating analysis". I leave it to readers to judge whether the passages I cite in the text ring true, or at least appear to have some basis in facts.
87. Barry (1942), pp. 161–2.
88. Ibid., p. 162.
89. Rakove (1979, p. 42) criticizes the (at that time) common view that at the First Continental Congress, Samuel Adams, John Adams, and Richard H. Lee "somehow manipulated events and debates to foreclose the possibility of reconciliation and enhance the likelihood of independence". He does not, however, address the issue of how the voting rules were adopted; in fact, he does not even mention the adoption of the rules on September 6, 1774. His strictures are mainly addressed toward those who, from Joseph Galloway onward, have asserted that the unanimous adoption by Congress of the Suffolk Resolves on September 17 was an effect of the "superior application" of Samuel Adams.
90. Jefferson (1950–) vol. 1, p. 324. As we shall see, Delaware adopted bound mandates in 1787 as well.
91. Jensen (1970), p. 132.
92. For the extent of the dilution, see Freedman (1992).
93. Jensen (1970), p. 148.
94. Jefferson (1950–), vol. 1, p. 327.

95. Ibid., p. 326. The last statement seems to assume, implausibly, that the *small* states have common interests that will enable them to act as a unitary pivotal actor. In a letter to John Adams from May 16 1777, Jefferson asked him to propose to Congress that "any proposition might be negatived by the representatives of a majority of the people of America, or of a majority of the colonies of America. The former secures the larger the latter the smaller colonies". The second claim also seems to rest on an assumption that the small states will have common interests.
96. Jensen (1970), p. 145.
97. Individual voting by members of equal-sized delegations would, however, have shown up the spurious nature of many "unanimous" decisions (Schlesinger 1968, p. 412).
98. For fuller discussions, see Jillson and Wilson (1994), pp. 138–42, 157–62; also Brant (1948), pp. 106–8.
99. Adams (1976b), p. 593.
100. Madison (1962–), vol. 6. p. 291; see also ibid., p. 310, for some modifications.
101. I have not seen evidence that state legislatures acted as either unconditional or conditional cooperators during the war. Unconditional cooperation in the war against Britain would imply that a state mobilized more when it could do more damage to the British rather than when itself was more exposed to danger (as predicted by the self-interest hypothesis). Conditional cooperation would imply that a state would mobilize more when it observed other states mobilizing rather than less (as that hypothesis would predict). After the war, the very notion of cooperation breaks down because, as (following Madison) I note below, the notion of the common interest was not well-defined.
102. McDonald (1979), p. 38.
103. Galloway (1780), p. 74.
104. Dougherty (2001), p. 3 and *passim*.

References

(*References are online)

Acemoglu, D., J. Robinson, and T. Verdier. 2004. Kleptocracy and Divide-and-Rule. *Journal of the European Economic Association* 2: 162–192. Papers and Proceedings.

Adams, J. 1976a. Diary. In *Letters of Delegates to Congress*, vol. 1. Washington, DC: Library of Congress.

———. 1976b. Notes on Debates. In *Letters of Delegates to Congress*, vol. 4. - Washington, DC: Library of Congress.

Ammerman, D. 1974. *In the Common Cause*. Charlottesville: University Press of Virginia.
Barry, R. 1942. *Mr. Rutledge of South Carolina*. Salem: Ayer.
Brant, I. 1948. *James Madison: The Nationalist*. New York: Bobbs-Merrill.
Breen, T. 2004. *The Marketplace of Revolution*. Oxford: Oxford University Press.
———. 2010. *American Insurgents, American Patriots*. New York: Hill and Wang.
Bumsted, J. 1974. 'Things in the Womb of Time': Ideas of American Independence 1633 to 1763. *William and Mary Quarterly* 31: 53–64.
Burnett, E. 1964. *The Continental Congress*. New York: Norton.
———. 1974. *The Continental Congress*. Westport: Greenwood Publishing.
de Montaigne, M. 1991. *Essays*. London: Allen Lane.
de Tocqueville, A. 2004. In *Oeuvres*, ed. Pléiade, vol. III. Paris: Gallimard.
Dougherty, K. 2001. *Collective Action Under the Articles of Confederation*. New York: Cambridge University Press.
———. 2003. Madison's Theory of Public Goods. In *James Madison*, ed. S. Kernell, 41–62. Stanford: Stanford University Press.
Elster, J. 1999. *Alchemies of the Mind*. Cambridge: Cambridge University Press.
———. 2012. Constitution-Making and Violence. *Journal of Legal Analysis* 4: 7–39.
———. 2015. *Explaining Social Behavior*. rev. ed. Cambridge: Cambridge University Press.
———. 2017. On Seeing and Being Seen. *Social Choice and Welfare* 49: 1–14.
———. forthcoming. The Political Psychology of Publius. In *The Cambridge Companion to the Federalist*, ed. J. Rakove and C. Sheehan. Cambridge: Cambridge University Press.
Franklin, B. 1754. Reasons and Motives for the Albany Plan of Union. *Papers, Yale University Press* 5: 399–417.
———. 1959–. *Papers**.
Freedman, E. 1992. Why Constitutional Lawyers and Historians Should Take a Fresh Look at the Emergence of the Constitution from the Confederation Period. *Tennessee Law Review* 60: 783–838.
Galloway, J. 1780. *Historical and Political Reflections*. London*.
Gould, C. 1986. The South Carolina and Continental Associations. *The South Carolina Historical Magazine* 87: 30–48.
Haw, J. 1997. *Edward and John Rutledge*. Athens: University of Georgia Press.
Hazard, E. 1794. *Historical Collection of State Papers*. Philadelphia: Dobson.
Hume, D. 1978. *A Treatise of Human Nature*, ed. L.A. Selby-Bigge. Oxford: Oxford University Press.
Hutson, J. 1987. Riddles of the Federal Constitutional Convention. *William and Mary Quarterly* 44: 411–423.
Irvin, B. 2003. Tar, Feathers, and the Enemies of American Liberties 1786–76. *New England Quarterly* 76: 97–138.
Jefferson, Thomas. 1950–. *Papers*. Princeton: Princeton University Press.

Jensen, M. 1950. *The Articles of Confederation*. Madison: University of Wisconsin Press.
———. 1968. *The Founding of a Nation*. New York: Oxford University Press.
———. 1970. *The New Nation*. New York: Vintage Books.
Jillson, C., and R. Wilson. 1994. *Congressional Dynamics: Structure, Coordination and Choice in the First American Congress 1774–1790*. Stanford: Stanford University Press.
Kromkowski, C. 2002. *Recreating the American Republic*. Cambridge: Cambridge University Press.
Madison, J. 1962–. *Papers*. Chicago: University of Chicago Press.
Maier, P. 1991. *From Resistance to Revolution*. New York: Norton 1972.
Mason, G. 1970. *Papers*. Chapel Hill: University of North Carolina Press.
McDonald, F. 1979. *E Pluribus Unum*. Indianapolis: Liberty Fund Press.
Miller, C. 1985. Social Development of the Colonial Chesapeake. *American Presbyterians* 63: 333–340.
Morgan, E. 2012. *The Birth of the Republic*. Chicago: University of Chicago Press.
Muehlbauer, M. 2008. Justice and Just War: A History of Early New England 1630–1655. PhD Dissertation, Department of History, Temple University*.
Nelson, J. 2011. *With Fire and Sword*. New York: Thomas Dunne Books.
Rakove, J. 1979. *The Beginning of National Politics*. Baltimore: Johns Hopkins University Press.
Ramsay, D. 1990. *The American Revolution*. Indianapolis: Liberty Fund Press*.
Schlesinger, A. 1918. *The Colonial Merchants and the American Revolution*. New York: Atheneum.
———. 1968. *The Colonial Merchants and the American Revolution*. New York: Atheneum*.
Shannon, T. 2000. *Indians and Colonists at the Crossroads of Empire: The Albany Congress of 1754*. Ithaca: Cornell University Press.
Simmel, G. 1908. *Soziologie*. Berlin*.
Smith, P. 1976. *Editorial Comments on Galloway's Proposal in Letters of Delegates to Congress*. Vol. 1. Washington, DC: Library of Congress.
Starr, R. 2000. Political Mobilization 1765–1775. In *A Companion to the American Revolution*, ed. J. Greene and J. Pole, 222–229. Oxford: Blackwell.
Stone, F. 1889. *History of the Celebration of the One Hundredth Anniversary of the Promulgation of the Constitution of the United States*. Philadelphia.
Ward, H. 1961. *The United Colonies of New England 1643–90*. New York: Vantage Press.
Weslager, C. 1976. *The Stamp Act Congress*. Newark: University of Delaware Press.
White, M. 1987. *Philosophy, The Federalist, and the Constitution*. New York/Oxford: Oxford University Press.
Wood, G. 1991. *The Radicalism of the American Revolution*. New York: Vintage Books.

A Political Theory of Constitutional Democracy: On Legitimacy of Constitutional Courts in Stable Liberal Democracies

Pasquale Pasquino

In this chapter I shall focus on the question of the legitimacy of European Constitutional Courts (hereafter referred to as CC.) I assume that everyone knows what these courts do. In focusing on their legitimacy, I want to analyze the question of the rational (I understand this term in the minimalist Hobbesian sense)[1] arguments we can present to support and justify to ourselves as citizens the existence of a CC in a constitutional democracy (*verfassungsmäßiger Rechtstaat*) (*stato di diritto costituzionale*).[2]

Before explaining what I've tried to do in this text, I need to say a few words about what I *do not*. Discussing a research project obviously demands checking the coherence, the "integrity" of the arguments presented, but it has also to be clear about precisely which question the author wants to ask and tries to answer, for it is no sound objection to say that she has failed to answer a question outside the intended scope of her research. The answer may be unclear or unpersuasive (in a strong, rigorous sense of the word, in a research like the one presented here, it cannot be simply true or false), but the question itself can only be unclear and perhaps uninteresting—which is a

P. Pasquino (✉)
Department of Politics, New York University, New York, NY, USA

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_8

subjective evaluation and depends mostly on what we can call "circles of recognition."

To begin with here, what I am not trying to explain and justify. I do not want in my research to talk about the role of constitutional courts in fragile or illiberal democracies, about American judicial review, or transnational/supranational courts.

Assuming the definition of *democracy* offered by Adam Przeworski in a number of articles (Przeworski 1999) (i.e., that a democratic regime is one in which the incumbent government can lose elections—so that Cuba or China do not come under this category), I can be even more specific. I will not consider the role, function, and legitimacy of the Supreme or constitutional court in countries like Azerbaijan, Georgia, Egypt, Turkey, or Pakistan, nor of the new CC of Latin America. I need however to add a supplementary qualification. The case of Turkey is particularly interesting. Since 1961, there is a constitutional court in Turkey which has been working pretty effectively (until recently) as guardian of the Kemalist constitutions. Turkey corresponds, by the way, to the minimalist criteria of a democratic regime according to Przeworski: the incumbent party lost the election not only of 1950, but more relevant, the Kemalist political elite was repeatedly defeated in the last 12 years, since the Islamic party AKP (*Justice and Development Party*) took power, without being successfully challenged by military intervention. So a rotation in power seems to be a reality in Turkey (provided that the AKP doesn't place obstacles to it in the future); the reason why I exclude this country from my analysis is that Turkey, so far, doesn't look like a liberal democracy (the treatment of Kurds and of sectors of the opposition in the country is well known and an evident example of disrespect for fundamental citizens' rights) (Zakaria 2003).

So the object of my inquiry is limited to stable liberal democracies (notably Germany, France and Italy), by which I mean those political systems that have constitutions resulting from the stable compromise between different social political groups who believe, in principle, in the same basic values, and accept the idea of *limited government*.

As a footnote, I would like to add that a comparative analysis of constitutional adjudication mechanisms should distinguish four basic subgroups of institutions: (1) the American type of Supreme Courts with competence of judicial review of primary legislation (for instance, the Supreme Courts of India and Japan); (2) the constitutional courts of European continental type (those I analyze in this research, but also, Poland, Spain, Portugal, etc.);

(3) the important family of Constitutional/Supreme Courts of quasi-democratic, semi-authoritarian or illiberal countries (like Turkey, Egypt, Tunisia, Russia); and (4) Courts which do not seem to do anything or just rubber-stamp the decisions of the executive (Georgia, Azerbaijan, Ivory Coast, etc.).

It is, moreover, important to draw attention to the circumstance that in speaking of constitutional courts, it is difficult to say anything from a normative/justificatory point of view if we do not first have a look at the specific constitution that the Court is supposed to protect and guarantee. As the case of Turkey shows, a constitutional court can quite effectively protect a constitution imposed by a tiny minority over a population that never freely accepted it. These types of radically transformative constitutions (of Jacobin type) are not the object of my research, even though I believe that they are of extraordinary political interest.

* * *

My intellectual enterprise is both descriptive and normative and normative in a sense that can be qualified as *justificatory* rather than *revisionary*.[3]

I have certainly a preference for the German model of constitutional adjudication vis-à-vis the American Judicial Review (a preference that has no significant importance—I have no transformative claim—but it inserts a normative dimension into my descriptive enterprise, since the Courts I discuss are one of the possible models of constitutional adjudication).[4] My goal, in any event, is primarily to claim that the existing institutional setting (the presence of a divided power of *Rechtserzeugung*—law-making power—between elected bodies and courts of justice) is the best form of government (in Churchill's sense of this ambitious expression) we have been able to establish, rather than assuming the posture of the reformer suggesting important, significant, and wonderful (and probably impossible) transformations of our institutional and constitutional order. So it is more a sort of apology for the status quo than plea for doing better in the countries of which I'm speaking in my work—even though it is possible to discuss minor reforms concerning the functioning of these institutions.

To be faithful to myself, I want to add that I have no hostility at all toward the idea of improving the status quo. Generally speaking I would say the contrary. All the societies in which we are living in the West are to different degrees fundamentally unjust, in my personal opinion.

But my maxim is that before trying to change the world, we need to understand it and to see also what the positive achievements are amidst the increasingly unjust conditions, both social and economic.

* * *

There Are No Rights Without Remedies

Constitutional democracies are political systems where non-elected, non-accountable organs (usually called courts) can modify through interpretation or simply cancel statutory legislation enacted by elected and accountable parliaments.[5]

With the authors of the *Encyclopedia Britannica*,[6] I believe that this political system is different from the one imagined by the authors of classical representative government, both in France and in the USA, or, to use the English expression, it is unlike the Westminster Model of government (with very weak, if any, judicial review/constitutional adjudication, depending on one's definitions of these terms).[7]

Some (few?) people seem to believe that this change vis-à-vis modern representative government is irrelevant or marginal[8] since these organs cannot do anything contrary to the will of political (=elected) actors. Elsewhere I have to discuss extensively this point, more exactly the latitude of the Constitutional Courts' discretionary power—which is not an *all or nothing*, but of the order of the *something*. Now I shall take issue with the large body of literature that has recognized this crucial transformation of representative government introduced by modern constitutional regimes, which establish an organ with competences, which are somehow different in different legal systems. I will focus here on the question of a constitutional court's *legitimacy*.

Without entering into a conceptual analysis of this term, I need to specify the sense in which I use this concept. The word legitimacy (starting from the seminal work of Max Weber) has a double meaning. From an empirical point of view, constitutional courts are among the institutions of contemporary democracies that have the best reputation among the citizens. (This is the case in countries like Germany, France, and Italy, less so in Spain, because of the tensions between the central government and the provinces of the young Spanish democracy, notably Cataluña, tensions that the Constituent Assembly decided to leave open and up to the *Tribunal Constitutional* to settle.)

This type of popular legitimacy or social approval[9] (one should consider that parliaments and political parties have lost such approval dramatically because they have more and more the reputation of narrow partiality/partisanship) is in itself very important, though not the specific object of my intellectual investigation. I am looking for the reasons/rational arguments that could support the existence of such an institution as the Constitutional Court.

One might note here that the classical twentieth-century theories of democracy do not discuss constitutional adjudication. Hans Kelsen speaks only en passant of it in his *Wesen und Wert der Demokratie* (1929) [translated as *The Essence and Value of Democracy* (2013)].[10] Less surprisingly, Schumpeter never refers to this aspect of a contemporary democracy in the famous chapters in his book *Capitalism, Socialism and Democracy* (1942), since what he had in mind when speaking of democracy was the British political system of the twentieth century.[11] The same absence persists in books on democracy by Giovanni Sartori (1987).[12]

Still, the question of the legitimacy of CC is anything but new: it was discussed not only in the US in the nineteenth century but also with an extraordinary richness of arguments and counterarguments by law professors during the Weimar Republic[13] and by political actors in Italy during the process of making the republican constitution in Rome in 1946–47 when the institution of a constitutional court was strongly opposed by the Socialist and Communist members of the Constituent Assembly (Pasquino 2006a).

It is interesting to notice that there is a remarkably repetitive character within these debates, for reasons partially connected with the fact that comparative constitutional theory has paid so far only a limited attention to these German and Italian arguments.

Criticisms of CCs often tend to revolve around the following points, discussed very well by Mauro Cappelletti in his seminal work *Giudici Legislatori?* (1984: 72–82):

(a) The difficulty ordinary citizens have in understanding the Constitutional Courts' opinions (the objection being that they are in some sense, aristocratic, here referring to the technical dimension of judge-made law, hence the difficulty of accountability, the precondition of which, in theory at least, being that the citizens understand what law-makers do);
(b) The occasional retroactive character of judicial decisions (contrasting with the principle of "no retroactive law"—*Rechtsicherheit*);

(c) The institutional ignorance of judges and its impact on law-making (often in relation to decisions which imply a large set of specific and nonlegal knowledge);
(d) The anti-majoritarian character of the judicial law-making.

In his book, Cappelletti rejects these criticisms with robust arguments.

Still, the main challenge to constitutional adjudication by courts of justice through a panoply of arguments (systematically repeated by a large number of critics) is that Constitutional Adjudication is *undemocratic*. Simply formulated, the claim boils down to the following point: if modern democracy is a governmental order, in which the exercise of political authority is based on a mechanism of popular authorization: elections, then those governmental organs that are not elected by the citizens and so not accountable to the voters are incompatible with representative government. One can think here of the open opposition of E. Sieyes to the royal veto in 1789: the king cannot be (co-)legislator, he said, since he is not elected and accountable (Pasquino 1998a)—but also of J. Madison's difficulty in justifying the fact that the members of the judiciary do not respect the "republican principle."

* * *

In an important book published in 1931, *Der Hüter der Verfassung* (The guardian of the constitution),[14] Carl Schmitt launched an upfront attack against Kelsen's text of 1928,[15] the first theoretical foundation of constitutional adjudication in Europe. In his book, the German constitutional lawyer did not reject the idea that a modern constitutional democracy, like the Weimar Republic, needs a guardian of the supremacy of the constitution. In fact, in his *Verfassungslehre* (published 3 years earlier, 1928), Schmitt, the theorist of the constituent power of the people, clearly endorsed the idea that the constitution is a political decision *superior* to statutory legislation enacted by an elected parliament, since this one cannot modify the constitution with the same procedures used to enact laws.[16] What Schmitt rebukes, on the basis of his democratic ideology (where legitimacy coincides with elections), is the Kelsenian doctrine that the guardianship, of what the Austrian colleague called a "hierarchy of norms" between the constitutional provisions and the statutory legislation, should be attributed to a judicial organ, meaning to a non-elected and non-accountable court of justice.

The first part of Schmitt's book is a vehement attack on Kelsen, aiming to show that the judiciary should not be allowed to exercise the function of guardian of the constitution for two primary reasons: (1) because a constitutional court cannot simply operate through judicial syllogisms, that is, the mechanism of subsumption of the statutory norm under the constitutional provision, which consists in merely checking the non-contradiction between the two norms of different level, to use the metaphor of the pyramid (the *Stufenbaulehre*),[17] as Kelsen seemed to claim, and (2) because this function of control is an eminently "political" one.[18]

Equally important is the last part of the book (Schmitt 1931: 204–242) where the German constitutionalist defends the idea that only a *democratic*, that is, a *popularly elected* organ, can assume the function of guardianship of the constitution: the President of the republic, elected and accountable, is the only agent then who can exercise this crucial function.

I do not need to discuss these issues here nor show the paralogism of Schmitt's theory of the "neutral power."[19] I want instead to stress that the accusation of incompatibility between constitutional adjudication exercised by a court of justice and democracy is not new, and that any theory of democracy which reduces this form of government to the electoral accountability of the governing organs has to repeat Schmitt's claim that the judges lack the legitimacy for important political decisions and tends toward denouncing such a constitutional court as an *aristocratic* institution.[20]

A possible counter to Schmitt's challenge has been to say that Courts that exercise a constitutional review of statutory legislation are not acting as legislators, and so are not usurping this function from its rightful (elected) organ. Thus, there is no reason for democrats (more exactly, *electoralists*, *électionists*[21]) to worry about and be critical of constitutional adjudication by a court of justice.

On this point, I side with the critics. The idea first presented by Kelsen under the label of *negative legislation* (Kelsen 1945: 267–269),[22] does not withstand any serious scrutiny. Likewise, the defense of the Court based on the suggestion that because often members of constitutional courts are appointed by elected representatives, they are, therefore, democratic (von Bruneck 1988: 224). The last argument sounds like sheer Hegelianism: everything is connected with everything. Such a pseudo-answer begs the question of legitimacy with a conceptual pirouette!

What I want to dispute is the criticism based on the ancient dogma of the separation of powers: specifically, the claim that constitutional adjudication represents an encroachment upon the legislative function. That seems a

bold claim: qualifying as *dogma* the pillar of modern liberal constitutionalism that goes under the etiquette of separation of powers. But I'm not an anti-liberal, nor a subversive. I am repeating a point made quite persuasively by Hans Kelsen (1945: 269).

Still since this claim, or more exactly *my own version of it*, plays a very important role in my entire argument, I need to clarify what I mean before proceeding.

I do believe that the Constitutional Courts (the European name for governmental organs which exercise constitutional adjudication but which are not elected and not accountable to the voters) do exercise a legislative function; they are, indeed, to use the expression of Michel Troper, *co-legislators*. However, I do not see in the Constitutional Courts' exercise of this function a form of despotism—on the contrary, their participation in the *law-making* function of the political authority seems a useful mechanism and one that may, in fact, create crucial obstacles to despotic, authoritarian, or illiberal governments[23] and help the stabilization of liberal democratic regimes by improving their ability to protect constitutional rights.

If we agree that interpreting and canceling statutory legislation (declared unconstitutional by the Constitutional Court) is *Rechtserzeugung* (we can translate the Kelsenian expression as "law-making"), it is not enough to claim, with Cappelletti, that it is a diverse form (different from the parliamentary *Rechtserzeugung*) of law-making, which is certainly true (Cappelletti 1984: 63). We have to explain why it is good. More specifically, we must grasp why it is good that law be produced by two different types of institutions, and in which sense they are different, even though they both make laws, that is, binding decisions for the members of the political community.

To do this we need to step back and have a look at the rationale of the Montesquieuian doctrine of the separation of powers.

Powers/Functions

Given that this point, that is, the doctrine of the separation of powers, is not always clearly understood,[24] we need to remember that, in the famous Chap. 6 of the book XI of his *The Spirit of the Laws* on the Constitution of England, Montesquieu distinguished (not always consistently) between *functions* exercised in each political system and *branches* or agencies/institutions exercising these functions.[25]

> In every government there are three sorts of power: the legislative; the executive in respect to things dependent on the law of nations; and the executive in regard to matters that depend on the civil law.[26]

The three functions in this first classification correspond to the taxonomy proposed by Locke in the *Second Treatise*: legislative, federative, executive, though in the next paragraph of the same chapter Montesquieu modifies the names of the tripartite classification:

> By virtue of the first, the prince or magistrate enacts *temporary or perpetual laws*, and amends or abrogates those that have been already enacted. By the second, he makes peace or war, sends or receives embassies, establishes the public security, and provides against invasions [this is evidently the federative power].[27] By the third, he punishes criminals, or determines the disputes that arise between individuals. The latter we shall call the *judiciary power* [italics mine], and the other simply the executive power of the state.[28]

Now the federative takes the name of executive *function* and the third function (to judge and punish) the name of judiciary. In the somewhat imaginary conception of the English constitution[29] presented by Montesquieu in this chapter, the judicial function (called also ambiguously *puissance* or *pouvoir*) is famously a "null power."[30] The federative/executive function is not really an object of discussion[31] and the analysis focuses on the legislative function, which is also (as in Bodin and Rousseau) the supreme (sovereign) function/power.

The crucial mechanism needed to avoid a despotic regime was for Montesquieu to stay away from any form of monocratic exercise of the *law-making function*:

> In such a [here the translation is not accurate, better: 'In a'] state there are always persons distinguished by their birth, riches, or honors: but were they to be confounded with the common people, and to have only the weight of a single vote like the rest, the common liberty would be their slavery, and they would have no interest in supporting it, as most of the popular resolutions would be against them. The share they have, therefore, in the legislature ought to be proportioned to their other advantages in the state; which happens only when they form a body that has a right to check the licentiousness of the people, as the people have a right to oppose any encroachment of theirs.

> The legislative power is therefore committed to the body of the nobles, and to that which represents the people, each having their assemblies and deliberations apart, each their separate views and interests.[32]

It is clear that to avoid despotism, Montesquieu, in speaking of England, presents a model of the mixed constitution[33]—here used to divide the sovereign legislative function—typical of the post Glorious Revolution English political order.

The judiciary in turn has to be independent[34] from the law-making function since it would be unacceptable that the judges (or those who exercise judicial function) apply the law arbitrarily. All this is written in the chapter on England. But in the fundamental chapter on legal government (*The Spirit of the Laws*, II, 4), Montesquieu developed the idea of a *dépôt des lois* which endowed the high courts of justice (the *Parlements d'ancien régime*) with some active role in the law-making function:

> It is not enough to have intermediate powers in a monarchy; there must be also a depositary of the laws. This depositary can only be the judges of the supreme courts of justice, who promulgate the new laws, and revive the obsolete.[35]

This distribution of the law-making function[36] between different and independent branches/agencies is for him as well as for modern constitutionalism the essential tenet of a liberal anti-despotic, anti-authoritarian form of government.

The contemporary version of the divided law-making power has been presented recently as the end of the democratic regime and as a simple revival of the pre-modern, mixed regime.[37] In a very useful article devoted to the separation of powers in Montesquieu, Michel Troper[38] wrote recently:

> It is paradoxically a variety of the balance of powers à la Montesquieu that best survives. Assuredly not as L'Esprit des lois describes it, in other words between a noble House, an elected House, and a king armed with a veto, but today we know another form of it. *In most countries the legislative power today is shared between parliamentary assemblies and constitutional courts.* And if one proposes several justifications for control of the constitutionality of laws, the most widespread and most effective is by far that which makes courts into counter-powers.[39] Obviously Montesquieu said nothing about constitutional courts, but this justification can claim his legacy for several reasons: it allows, it is said,

> preservation of political freedom conceived as submission to the law, understood in a broad sense, in other words as submission to the constitution; it consists in arranging for power to check power; and it allows bringing the control of constitutionality under a form of *mixed government*, since the will of the parliamentary majority of the moment, the democratic element, is controlled by a court composed of persons chosen for their competence, thus by an aristocratic element. The difficulty confronted by tenants (sic) of this justification is, however, to assume Montesquieu's heritage entirely on two principal points: the conception of the power to judge as nil and thus unable of playing a role in the balance of powers, and the conscious acceptance of mixed government and the *correlative rejection of democracy*.

Troper refers here to the idea that the role of the constitutional courts in the contemporary political system we call *democracy* represents a revival of the doctrine of the mixed constitution.[40] However, he draws an unclear conclusion from it: that we have to forego calling our systems democracies and, instead, accept explicitly the mixed government. I need to discuss this claim since it presents as accepted evidence something that is based on implicit and disputable assumptions.

Democracy, in Troper's language, seems to be the form of government where law is the will of the representative (parliamentary) majority. In the classical theory of the forms of government, such a regime would have been qualified not as democratic, but at best, as an *elective oligarchy*, democracy being the self-government of the *demos* (in the original sense of the Greek term,[41] and in the Aristotelian tradition the word demos had the meaning of middle-lower classes, the best modern translation of the Greek *aporoi*). In modern political language, the democracy of Troper is a representative government, such as it exists notably in the UK but not anymore in the very large and constantly increasing number (Ginsburg 2003)[42] of countries which, after the Second World War, introduced, in different waves, constitutional courts. Now, since it is evident that constitutional democracy is not the same regime as the representative elected oligarchy established in France at the end of the eighteenth century, we need to clarify in which sense the new regime—that is, constitutional democracy—is a mixed government. More specifically, we have to see if it is a real equivalent of the *memigmene politeia* (the mixed government) of the Aristotelian, Polybian, and Machiavellian type, the same that we find revived later in the polyarchic structure of the divided legislature of which Montesquieu speaks in his famous chapter on the Constitution of England.

The divided power, which *The Spirit of the Laws* suggested as an alternative to the French absolute monarchy, was based on the classical *anatomy of the city* of Greek origin according to which the political body is divided into substantive non-homogeneous parts (the Aristotelian/Machiavellian *mere tes poleos*, the parts of the city) which have different rights (to use our language) and must share political authority by participating actively in the government of the society (see the sections of Aristotle's Politics concerning the *memigmene politeia*, the section of Polybius' *Histories*, Book VI, devoted to the *Romaion politeia*, and Machiavelli's project of a constitution for Florence: *Discursus Florentinarum Rerum*, 1522) (Machiavelli 1989).

Now, if we use the same expression "mixed government" without specifying what we are speaking of, we run the risk of saying nothing conceptually useful.

Contemporary constitutional democracy is based on a legally equalitarian anatomy of the city where fundamental rights are the same for everyone: the abolition of aristocratic privileges is the common element of both the American and the French eighteenth-century constitutional revolution. Judges have no special rights; they are (supposed to be) experts, likewise the older Athenian citizens who manned the people's courts in the fourth-century BC (FN concerning the age of the *dikastai* and the less complex expertise required in a society like the one of the ancient *Demokratia*). If we want to speak of mixed government to qualify the political system, the *Encyclopedia Britannica* calls democracy (3)—I prefer to speak of *divided power*—we have to specify that government in this case refers to the legislative, or better, and the "normative" function/power, consisting in enacting law, that is, binding decisions for the entire community. Moreover, the *enemy* and the *antonym* of the ancient mixed regime were both oligarchy and democracy (see Machiavelli's *Discorsi*, Book I, Chap. 2), meaning the domination of one part of the city over the other. The enemy or the threat to the contemporary notion of divided power is governmental absolutism, or unlimited/arbitrary state power.

The apparent paradox of modern political theory is that absolutism was born to protect natural rights and established, indeed, the basis of liberalism, meaning here limited government (as Leo Strauss rightly stressed to his dismay (Strauss 1995), which I emphatically do not share at all!).

Now, what are the parts of the new mix and what are their specific qualities? We need to be clear about this point before asking what is good about the new mixed government.

In the classical doctrine, the *gnorimoi/euporoi* were, so to speak, ontologically (by nature) different kinds of people having superior qualities (Aristotle)[43] or justified (insuppressible) superior *humors* ("il desiderio di dominare") and interests (Machiavelli).

In the Hobbesian *society without qualities*, instead, people can be different only because of their knowledge (there are professors at NYU and people cleaning the apartments of those professors, who had no chance to go to good universities) with the same formal rights and the same dignity (at least in theory). In what sense, then, might the judges be different or superior? In what sense are they an aristocracy—to repeat the term used, somehow in the sense of the French word *nobility* of the Ancien Regime, to disqualify them and their function by radical democrats like Schmitt, Troper, and Waldron? They are superior in no substantive sense at all; their function is important and their expertise hopefully high, but they do not need to come from a particular social class (consider in the USA the cases of Samuel Alito[44] and Sonia Sotomayor[45] of the USSC) and have no special rights but only a constitutional function. Calling these people an aristocracy is sheer rhetoric[46] or populist anti-elitism, or more simply, it means that they are not electorally accountable, which by the way, is true and good for reasons I'll try to explain.

If our constitutional democracies are characterized by the division of the normative (*vulgo* legislative) function between electorally accountable organs and non-accountable ones, this is not because one social part can oppress the other,[47] but because even an elected and accountable majority can represent a threat to individual rights. The nature of the danger being different, different too is the nature of the remedy, and different also is the anatomy of the city which is presupposed in this new form of mixed government. The Schmittian defenders of democracy[48] will insist that the judges, not being electorally accountable, are therefore not democratic, where democracy boils down to elections, and elections are another name and, in fact, the only name for political legitimacy.

This is precisely the point I want to discuss and reject.

It is a fact that elections are not the only source of political authority in our constitutional democracy. Both judges and members of central banks are not electorally accountable, and a revisionary theory supporting the abolishment or the reduction of power for these authorities is conceivable (even if probably utopist). My project is to offer a defense of the existing political system, at least in countries like Germany, France and Italy.

Elsewhere (Pasquino 2013a, b) I have discussed the weakness of representative government based on simple competitive elections: rational, elected officials tend to be partial (to their voters) and myopic (looking always at the temporally limited horizon of their reelection).

Here I have been claiming, with the critics of constitutional adjudication, that the members of the Constitutional Courts are indeed co-legislators. I can even say that they exercise on top of this legislative function some incremental form of constituent power—constitutions being incomplete contracts,[49] it is de facto inevitable that, notably in case of *Organstreit*,[50] the Court has to "write" a fragment of the silent constitution. The Constitutional Courts, I wish to claim, have a significant binding power on the citizens—I shall discuss in the next section the limits of this power.

What is good about this power, and why is it rational to accept it, meaning the structure of the constitutionally divided normative power? This is the question to which I want to turn now.

A quotation from Hans Kelsen on democracy and liberalism can usefully introduce my argument:

> The transformation in the concept of freedom, from the notion of the individual's freedom from state rule to the notion of the individual's participation in state rule, also signifies democracy's detachment from liberalism. Because the demand for democracy is considered met to the extent that those subject to the state order participate in its creation, the ideal of democracy is independent of the extent to which the state order affects the individuals who create it – that is, independent of the degree to which it interferes with their "freedom." As long as state authority emanates from the individuals subject to it, democracy is possible even in the case of unlimited expansion of the state order over the individual – that is, complete annihilation of individual "freedom" and negation of the liberal ideal. And history shows that democratic state authority does not tend less towards expansion than autocratic state authority. (Kelsen 2013: 88)

That majoritarian democracy needs some correctives has always been a tenet characterizing modern constitutionalism. Article 16 of the French declaration of human rights reads:

> Any society, in which no *provision is made for guaranteeing rights* or for the separation of powers, has no Constitution.[51]

And the First Amendment to the American constitution starts with the words:

> Congress shall make no law respecting...

The name of these limits to the power of the elected majority is *fundamental rights*. They come from a cultural tradition older than the representative government of Madison or Sieyes (now called democracy): the theory of the modern state and more specifically the justification for the *political obligation* of the citizens to obey the commandments of the political authority.

The *contractarian* political philosophy of the seventeenth century laid the foundations of an important conception of political obligation, and it is this one I need to take into account in order to present what I consider the best possible justification for constitutional adjudication by courts of justice.

It has been claimed that this intellectual tradition supposes a contract at the origin of political authority.[52] The word *origin* is ambiguous and actually misleading. Neither Thomas Hobbes nor John Locke ever tried to present a doctrine concerning the historical origin of political power (this is for Hobbes, mostly originating in a conquest—*acquisition* in his language) (Hobbes 1651: Chap. XX) and, for Locke, in the slow transformation of a patriarchal structure of political authority into a limited form of government) (Pasquino 1998b). The object of the intellectual enterprise of the classical contractarians—and I'll focus here on Hobbes—was to offer an argument in favor of political obligation with, in other words, an argument concerning the reasons why it is rational—in their own interest—for citizens to obey the political authority, Leviathan, and *under which specific conditions*.

This point is often disregarded, but it is perfectly clear that for Hobbes it is rational and not self-defeating for the citizens to obey political authority *if and only if* it performs a function[53] which consists in the guarantee of the fundamental (Hobbes says *natural*) right of self-preservation of the subjects, *omnes et singulatim*. The obedience is not unconditional but presupposes an exchange between protection of the fundamental right to "life and limbs" of the members of the political body and their obedience.[54] As Spinoza will repeat later on, *oboedientia facit imperantem* so that political obligation to obey is at the same time the origin of political order (the commonwealth by opposition to the state of nature) and of the legitimacy of political authority, if some conditions, the protection of rights, are satisfied.

The concept of the *contract* doesn't refer for Hobbes to a real or tacit transaction; it is a mental experiment.[55] I want to suggest a similar thought experiment as a rational justification of constitutional adjudication.

* * *

> Democracy as a system of majority decision-making presupposes agreement on that which cannot be voted upon. (Arndt 1976: 128)

* * *

Suppose we are citizens of a representative government, more specifically of a Schumpeterian democracy, one in which the incumbent government, chosen through competitive elections, can and quite regularly loses the election to the challenger. I will ask you to decide under a thin *veil of ignorance* what structure of government you would chose. My veil of ignorance is *thin*, indeed, since I'm asking you simply to assume—which is normally the case in a stable competitive democracy—that you do not know whether the party or the coalition of the parties you prefer will be the majority or the opposition after election day and those for years after. You are also interested, I assume, in the protection of your fundamental (constitutionalized)[56] rights since your authorization (through the election) of political authority (pro tempore, until the end of the electoral mandate) is not a total alienation of your rights. You may rationally fear that the majority will not protect them, even more if it is not the one that you would have preferred that wins the election.

Would it not be rational in this situation to establish next to an elected and accountable parliament a court of justice that can possibly protect your fundamental rights?

Let's look more closely at this problem.

After the inception of the contractarian doctrine, born in the middle of the religious civil war, our Western conception of fundamental rights expanded from the simple guarantee of "life and limbs" to a larger number of positive and negative freedoms that we citizens do not want to see infringed upon by the government, even if it is an elected and accountable one.[57] When modern representative government was established toward the end of the eighteenth century, the agreement, so to speak, was not only: you citizens vote for us, representatives, and we will command you, with the clause that you citizens choose the government and can dismiss it after a given lapse of time; it also implied the promise of a guarantee of the constitutionalized political and civil rights, as I said referring to the

declarations of human rights. Abandoning these rights would be a sort of suicidal pact; in the Hobbesian logic, it would be perfectly irrational to obey the government.

Now, if we look at the *list of rights* that the political authority promises us to protect and guarantee, it is clear that many of them have a vague character (think of freedom and equality); there is actually a lot of indeterminacy concerning their content. This fact may and will *inevitably* produce conflicts of interpretation concerning the content of rights that, moreover, are never absolute. I want to maintain that this is perfectly physiological; we do not need to think that the majority is willing to establish a dictatorial power. I'm not assuming that we are in Hungary, where the young liberal democracy seems about to collapse,[58] but in stable democracies like France or Germany. I'm just surmising that the citizens and the government (the majority) may disagree as to the respect for the fundamental rights by primary legislation.

Disagreement on such questions is inevitable and even sound, since there is no true solution for such disagreements. So the question is, what to do in the case of such conflicts—conflicts of interpretation as to the content of the rights that the statutes, enacted by the elected legislative majority, have to respect? And I remember that in the liberal tradition, if the commandments of the political authority violate our fundamental rights, we are unbound from any political obligation to obey.

To allow that the citizens *uti singuli* be the judges of this conflict of interpretation would be a sort of recipe for anarchy (Hobbes and Locke spoke of *state of nature*). The radical democrats say that it has to be an elected organ, so the majority of the citizens (if we assume, which is normally wrong, that the majority of the representatives are an expression of the majority of the members of a political community). That would be another way of saying that "since we are politically majoritarian, we are legally/constitutionally right!"[59] It seems difficult to defend the thesis that the majority of the representatives are necessarily right in their interpretation of the fundamental rights.[60] This would be another way of speaking of a total abdication of rights in favor of a political elite. True, the abdication would be pro tempore and not *ad eternum*. Still, democracy might end up being, as Kelsen warned, a sort of despotism by rotation, each majority being able in turn to abuse citizens' rights, having the monopoly of their constitutional interpretation.

Having the possibility to appeal to an independent court of justice to adjudicate a conflict of interpretation between citizens and the public

authority seems prima facie a sensible and rational solution. Still, the democrat will object: Why should we trust the Court more than the elected and accountable majority? The answer is exactly: since the members of the constitutional court are *not* electorally accountable! Were they accountable, their function will be in a sense redundant, at least in a *Parteienstaat*—like Germany, French, or Italy; it would be just a third chamber, accountable to the same voters and likely with the same majority (or, then, with a different majority paralyzing the legislative process). Their independence and the obligation they have to justify their decisions (Cohen and Pasquino 2013) are reasons that should push us, under the veil of ignorance that I suggested, to opt for such an instrument in our choice of the set of institutions that is rational for us to choose. Notice that by *independence* I mean exactly non-accountability. The Court is independent since it has no reason to please the plaintiff or the government. The mandate of the judges not being renewable, they have no particular *incentive* to be biased in favor of one or the other party in the constitutional interpretation conflict. Their opinions (considering that in general the vote of the members of the CCs I'm considering is undisclosed) (Pasquino 2015) will not have any impact on the renewal of their office (which is impossible) or on possible other appointments at the end of their mandate.

It is evidently true that the establishment of a constitutional court creates a powerful organ[61] that is not under control of the voters, but this is the only way I know to establish a possible effective guarantee for our constitutional rights and avoid the pro tempore total alienation of them.

Three points deserve closer consideration:

1. The limits of the power of the CC—the absence of electoral accountability is not omnipotence.
2. This accountability limits the power of the elected majorities only if the next majority agrees with the citizen who claims that her rights are violated.
3. The mechanism of appointment of the members of the Constitutional Courts is an important element worth a specific discussion.

I'll discuss the first point at length in another article. Here I shall discuss briefly points 2 and 3.

The democratic argument in favor of political control over the elected branches (and here political means that control over decisions of the elected officials is exercised by the same elected official) boils down to the following

thought: Suppose the citizen Lambda believes honestly that a given statute or piece of primary legislation infringes upon her constitutionally protected rights. Suppose, moreover, that there is no court of justice where she can bring her case. The democrat will answer *either* that she is wrong since the representatives are right (so there is no possible agency problem—by *synecdoche*, the representative and the people are identical, going back from this point of view from Locke to Hobbes who claimed that there is no people without its representative, so no agency problem!) *or* that the legislator may be wrong but the only way to check whether the law is unconstitutional (meaning, violating rights that the constitution is supposed to protect) is to coalesce a new majority around the interpretation of the person who complains of the rights' interpretation by the government of her constitutional rights. One could say: Good luck! since plainly this last path is extremely difficult and de facto impossible for insulated groups which have no pivotal position in the democratic competition.

We see that in this vision, there is a single authorized interpreter of the content of the constitutional rights, the elite who wins the elections, and that there is no hierarchy of norms since the legislator alone is *at the same time* the author and the only interpreter of the law.

The argument against constitutional adjudication turns on the magic power of electoral accountability, by which in any event right (*le droit*) lies always where the number or strength (measured by ballots, once each 4/5 years) is.

The democrats may reply that it is even worse to give the *last word* concerning the content of fundamental rights to 9 or 15 non-elected, non-accountable judges who can impose arbitrarily (without any control) their will over the citizens and the elected representatives.

However, if we consider the mechanism of appointment of the judges sitting on the Constitutional Courts and how they make decisions, we may again find it rational to accept the structure of constitutional democracy and that we have an interest in sustaining it (rather than asking for the dismantling of constitutional adjudication).

I said earlier that elected officials have a special incentive to be partial to their voters, that this is the price we pay to have them accountable not to the voters in general but at least to their plurality (it is, indeed, the largest minority that in general produces a majority of representatives in the elected legislative assemblies). We cannot have neutrality and accountability at the same time, and by neutrality I mean the absence of the structural partiality

connected with the electoral mandate. This is a reason why it may be convenient to divide what I called the normative power/function between an elected and a non-elected organ. Moreover, we can decide, under the veil of ignorance, to establish a particular rule for appointing the members of the Court: a bipartisan mechanism that puts on the courts not only legal experts but also candidates who are accepted by both sections that normally struggle for governmental positions through competitive elections. This simple rule should produce a panel court where the members rather than opposing each other from radically diverse positions (which may happen when they are appointed by a simple majority) are more easily prone (than any ordinary elected assembly) to compromise and to function as an *intermediary body*—to use Montesquieu's expression that Alexander Hamilton repeated in *Federalist* #78[62]—between the elected representative majority and the citizen asking for protection of her rights.

This mechanism of appointment exists in Germany and for part of the members of the Italian constitutional court, and in my opinion is the one that we should choose under the veil of ignorance under which I suggested we run our mental experiment.

* * *

Citizens who are particularly *risk prone* or have very good (historical) reasons to trust their politicians (and to distrust legal experts) may think that they do not need this guarantee of their rights. They may want to accept that the government/the majority are at the same time the author of the laws and the judge of their constitutionality (of primary legislation's respect for citizens' fundamental rights). This is, though, not what many British citizens thought when they sent their complaints to Strasbourg to the *European Court of Human Rights* to ask protection of their rights vis-à-vis the British Parliament.

Jürgen Habermas in a debate with Ronald Dworkin years ago at the Cardozo Law School suggested that constitutional courts are needed in democratic societies *only* in countries like Germany because of its authoritarian past. I'm struck by the unusually parochial Habermas opinion. With very few exceptions, all democratic regimes in the world are either post-authoritarian or post-colonial; it is, in fact, only in the UK and among members of the British Commonwealth where the British were the absolute majority that contemporary democracy did not emerge as a post-authoritarian regime.

* * *

Most of the criticisms of constitutional adjudication concern the USSC,[63] which is a special court of justice, very old compared with the European Constitutional Courts I'm describing and defending, and characterized by life appointment (not easily compatible in my opinion with a republican culture, the one which doesn't recognize life positions for law-making organs) and by a mechanism of appointment of Justices that, in the absence of "divided government," often select judges with strong ideological biases. It may also be true that the USSC is too powerful or "activist"—from choosing the President of the Union (Bush v. Gore) to deciding on questions like same-sex marriage, something that has evidently to do with the will of the political branches not to take the electoral risk of some of these decisions. But I have neither the competence nor the intention to discuss these topics, which are the subject of entire libraries.

I believe that the only sensible way of discussing the legitimacy of the Constitutional Courts of the European type which I'm considering is to think of the institution as such rather than of its decisions from a liberal point of view, so that if a decision of the Court doesn't line up or side with our political preference, we believe ipso facto that this is a reason for its lack of legitimacy. The Italian parliament has been controlled for almost 20 years by a majority I deeply dislike. I never thought that this unhappy circumstance disqualified the institution of parliament. Nor would the fact that George W. Bush was the President of the USA for 8 years during which he made disastrous decisions for the country disqualified the US presidency. Courts most of the time make decisions that some people like and some others do not. This is simply inevitable. In general, the losers (either the citizen or the government) do not really like the decision that makes them losers—notably since they were no losers before the decision.

Our judgment on the Court's decision is certainly not more neutral than the decision of the members of such bodies. There is no objectively correct decision of a constitutional conflict. If there were, we should have been able to find out the mechanism producing this truth. Certainly, the solution, which consists in abolishing the possibility of such a conflict, as in China, doesn't look very appealing to me. Actually, in a pluralistic society, there is no such truth since democratic constitutional regimes are systems of limited authority and not institutions like the Catholic Church. Since we disagree about the content of many of our fundamental rights and understandably so, we may want to have a double check on the first interpretation, the one of the elected majority.

Is it then true that the Constitutional Court is the new sovereign and an absolute one? I do not think so. Two sets of consideration can be suggested here. The first one has to do with the so-called "last word" of the Court's decisions (Thibaudeau), the second with the limits of its discretionary power.

In the first, a radical attack on the idea of control over the decisions of an elected accountable parliament by a non-accountable body, the *jurie constitutionnaire* (Goldoni 2009, 2012) proposed by E. Sieyes in the year III, the member of the *Convention* Thibaudeau asked the old question of *quis custodiet custodes* to justify that elected officials had no need of another check than the popular one. Thibaudeau, and even more those who repeated his argument, conflates the final word in a litigation with the sovereign decision. The sovereign decision is the one which is not revisable, the sovereign being the agent who says: sic *volo* sic *jubeo stat pro ratione voluntas* (Thus I will, thus I command, my pleasure stands for a reason). Now, in a legal conflict, closure—the impossibility of a further appeal—is inevitable since in its absence the conflict would never terminate and the parties engaged in the litigation would have no hope of a solution and so no reason to enter into the legal dispute. The entire legal system of conflict resolution, in the absence of a final decision, would simply collapse and lose its *raison d'être*. From this point of view, the last word is simply unavoidable. But the solution of a legal dispute doesn't represent the end of debate inside the political system. Famously there are many cases in which the decisions of the Supreme/Constitutional Courts were neither accepted nor enforced by the other branches of the government, from those of the Marshall Court that President Jackson refused to enforce[64] to the more recent ones of the Italian Constitutional Court concerning the obligation of a pluralistic structure of the media.[65] Courts can speak (and write) decisions, but they cannot control and guarantee their enforcement. A constitutional decision is never the last world in a system of divided power.

It is a "second opinion" (Vermuele 2011) legally binding, but open to any sort of resistance (see the crucifixes in Bavaria after and notwithstanding the decision of the *Bundesverfassungsgericht*[66]) and revision and not only as often repeated by the constituent power.[67]

Divided power is, politically speaking, an open-ended decision-making system, where three actors play an equally important role: the voters, the elected representatives, and the Constitutional Court. Now, it is the existence of these three actors that can help us understand the limits of discretionary power of constitutional adjudication.

If we look simply at the legal dimension—from the point of view of a pure theory of law that refuses to take into account sociopolitical reality—

one could say that a constitutional court can interpret with extreme freedom general constitutional values or principles like "freedom" or "equality." This tells us essentially that the pure theory of law is blind, or at best one-eyed. In fact, and the fact in question implies the existence of a sociopolitical context in which each constitutional court happens to work, the members of such a collegial body not only have to persuade the collective body making the decision but also have to take into consideration (or at least they will be better off doing so) the preferences of the other major actors of the political system.

Graphs may help to make this final point (Fig. 1):

The distance between the preferences (the small circles) of these three main actors represents what we can call the latitude of *discretionary* power of the Court's decisions. It can choose any point inside the perimeter, which is not in its power to establish but which is imposed upon it. If, alternatively, the preferences of the three actors are closer, the Court has less significant room for making its decision. Here, again, the judicial body is constrained by forces that are out of its control.

Examples[68] are numerous. As to Fig. 2, we can think of *Korematsu* (v. United States, 323 U.S. 214, 1944) when President Roosevelt, the Congress, a large part of public opinion, and the Governor of California, Earl Warren, were in favor of the emergency measures taken against the American Japanese. Or of *Carolene* (United States v. Carolene Products Company, 304 U.S. 144) (1938) reversing Lochner, when Roosevelt, the Congress, and public opinion were on the same position. A similar argument may be made, in my opinion, concerning *Plessy v. Ferguson* 163 U.S. 537 (1896).

A clear example of Fig. 1 is *Youngstown Sheet & Tube Co. v. Sawyer*, 343 U.S. 579 (1952), when the USSC was able to oppose President Truman, who lacked the support of Congress and of public opinion.

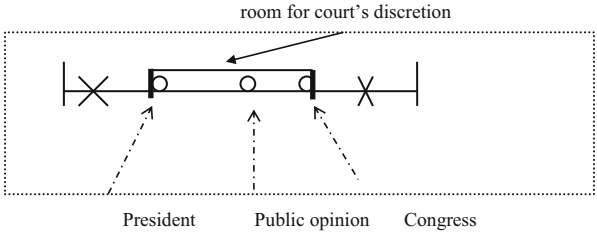


Fig. 1 Political possibilities I

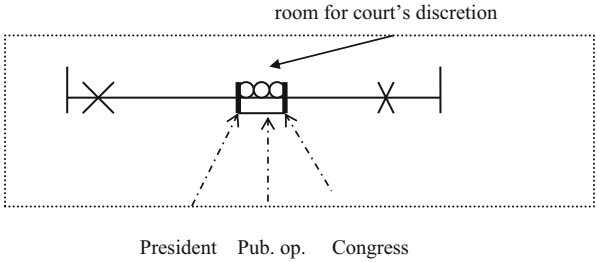


Fig. 2 Political possibilities II

Even the most famous decision of the French Constitutional Council—the one called the *bloc de constitutionnalité*, 1971—would not be understandable without the strong split between the *Assemblée Nationale* and the *Sénat* on the loi Marcellin.

Notes

1. I agree with Sharon Lloyd's interpretation when she writes, "Hobbes sought to discover rational principles for the construction of a civil polity that would not be subject to destruction from within [. . .] Hobbes further assumes as a principle of practical rationality, that people should adopt what they see to be the necessary means to their most important ends [notably the natural right of self preservation]" (Lloyd 2014).
2. On the meaning of this expression, see Pasquino (2012).
3. I use this term following D. Parfit (1986). The term *revisionary* was introduced by P.F. Strawson (1959) speaking of different types of metaphysics.
4. I'm now old enough to know the very limited possibility of modifying entrenched constitutional conventions, and more important, I have no pretensions at all to suggest anything to my American colleagues being a sort of institutional pluralist and not an expert on the American political and constitutional system. By "institutional pluralist" I mean, in the tradition of Machiavelli's *Discorsi*, that what is good and possible for Florence may not be possible for Naples. Or, to use a more contemporary example, that good institutions for Sweden (uni-cameralism and parliamentary system with an incipient and timid constitutional adjudication) are not exactly the same as those that are good for the USA or for Afghanistan.
5. I define this form of government by the presence of three elements: (1) a representative government based on universal suffrage, where there are regular, repeated, and competitive elections; (2) a *rigid constitution*,

encompassing fundamental rights and some form of separation in the exercise of political authority; (3) an independent judicial organ in charge of the guardianship of the constitution, which is called in Europe a constitutional court, council, or tribunal.

By accountability I mean the need of an agent or organ elected pro tempore to return to the electoral body to be renewed in her/his mandate. There are many other possible definitions, but in my text the term means only and exclusively what I stipulate.

A definition of constitutionalism as a key aspect of constitutional democracies is the one offered by J. Weiler (2011: 9) that I share: in a constitutional legal order "the constitution meant a higher law with the apparatus of judicial review and constitutional enforcement."

6. Sub voce democracy we read: "(3) a form of government, usually a representative democracy, in which the powers of the majority are exercised within a framework of constitutional restraints designed to guarantee all citizens the enjoyment of certain individual or collective rights, such as freedom of speech and religion, known as liberal, or constitutional, democracy" *The New Encyclopedia Britannica* (1993); p. 5 sub voce Democracy, that the guarantee of rights is for good reasons the task of a court of justice (rather than other possible alternatives) is the object of this text.
7. On this question, see now the important book by Gardbaum (2013).
8. This seems to be the opinion of Adam Przeworski (2011) where p. 160 the author shows, moreover, strong skepticism vis-à-vis constitutional adjudication and a clear preference for *majoritarian* democracy (on this book see my review in *La vie des idées*: http://www.laviedesidees.fr/Le-peuple-en-democratie.html?lang=fr)
9. Concerning the German Constitutional Court, see: Simon (1994).
10. See a partial English translation of this book in: http://publishing.cdlib.org/ucpressebooks/view?docId=kt209nc4v2&chunk.id=ch01&toc.depth=1&toc.id=ch01&brand=ucpress Now there is a complete translation of this fundamental text: (Kelsen 2013).

    In the Chapter on Administration, there is the only reference to constitutional adjudication (p. 83): "...not only individual administrative acts but also general regulative norms and especially laws can and must be submitted to judicial control – the former with respect to their legality and the latter with respect to their constitutionality. This control falls under the jurisdiction of a constitutional court, whose function is all the more important for democracy, the more the enforcement of the constitution in the legislative process is in the eminent interest of the minority and the more the rules regarding quorum, a qualified majority, etc., serve – as we have already seen [in the chapter I] – to protect that minority. [...] The fate of modern democracy depends to large extent on a systematic development of all

types of institutional controls. Democracy without such controls is impossible in the long run."

11. In the crucial section of his book devoted to the conditions for the success of the democratic method, the Austrian economist wrote nonetheless this important remark—on which I have to come back in another section of my research:

    "The second condition for the success of democracy is that the effective range of political decision should not be extended too far. How far it can be extended depends not only on the general limitations of the democratic method which follow from the analysis presented in the preceding section but also on the particular circumstances of each individual case." p. 291.

    (http://sergioberumen.files.wordpress.com/2010/08/schumpeter-joseph-a-capitalism-socialism-and-democracy.pdf)
12. An important exception is the American political theorist Robert Dahl (1989).
13. See some of the most relevant texts of this debate in Vinx (2015).
14. See now the partial English translation of this text in the book quoted at the FN 32.
15. See FN 48 hereafter.
16. Schmitt distinguishes not only statutory legislation from the constitutional provisions (*Verfassungsgesetzte*) but also the latter from the *positive Verfassung*, the constitutional core, which can be modified only by the citizens, the holders of the *pouvoir constituant*. This point has been repeated by the German Constitutional Court that claimed that only the German people and not even the elected representatives can abandon the German national sovereignty in favor of an European federal state (what is the real core of the German national sovereignty or identity (?) is not clear either in Schmitt or in the famous *Lissabon Urteil* of the German *Bundesverfassungsgericht*).
17. Elsewhere (Pasquino 1994a), I argued that Schmitt and Kelsen were not really speaking of the same question. Here I'm simply trying to show that the objections that Schmitt presented against the constitutional adjudication are a sort of *Ur*-criticism later on systematically repeated.
18. I discuss the sense of word "political" in this context in QUADERNI COSTITUZIONALI 2015.
19. See my text on the neutrality of Constitutional Courts (unpublished).
20. Schmitt's position on this question is presented accurately by Le Divellec (2007).
21. Emmanuel Sieyes used the word *électionisme* in his manuscripts to characterize his doctrine of the representative government. Electoralism is used

here to qualify the theories of contemporary democracy that reduce this form of government to electoral accountability and majority rule.

22. See also Kelsen (1928); this text is the French translation (probably by Charles Eisenmann) of the text presented by Kelsen in 1927 at the meeting of the German-speaking professors of public law in Vienna.
23. One can think that the authoritarian regime established by *Fidesz* in Hungary through a constitutional revision deprived the Hungarian Constitutional Court of almost any power of controlling the government.
24. For a correct interpretation of the doctrine, see Manin (1989: 728) and the seminal articles on the same question by Charles Eisenmann (2002).
25. This distinction is already in John Locke's Second Treatise; see Pasquino (1998b).
26. Nugent translation [http://ia700305.us.archive.org/31/items/spiritoflaws01montuoft/spiritoflaws01montuoft.pdf], p. 151; the original text reads: « Il y a dans chaque État trois sortes de pouvoirs: la puissance législative, la puissance exécutrice des choses qui dépendent du droit des gens, et la puissance exécutrice de celles qui dépendent du droit civil. »
27. On the federative power and its modern developments, see Kaufmann (1909).
28. « Par la première, le prince ou le magistrat fait des lois pour un temps ou pour toujours, et corrige ou abroge celles qui sont faites. Par la seconde, il fait la paix ou la guerre, envoie ou reçoit des ambassades, établit la sûreté, prévient les invasions. Par la troisième, il punit les crimes, ou juge les différends des particuliers. On appellera cette dernière la puissance de juger, et l'autre simplement la puissance exécutrice de l'État. », ibidem.
29. On this fundamental: Landi (1981).
30. This expression means, in the best interpretation I know, not that the judicial function is without any power, but that it is not attributed to a permanent body of magistrates, since exercised by jurors: "The judiciary power ought not to be given to a standing senate" (Engl. Transl., p. 153).
31. On the executive function, I should refer to a few very important works: Necker (1792), Barthélemy (1906), Smend (1923), and Cheli (1961).
32. P. 155; "Il y a toujours dans un État des gens distingués par la naissance les richesses ou les honneurs; mais s'ils étaient confondus parmi le peuple, et s'ils n'y avaient qu'une voix comme les autres, la liberté commune serait leur esclavage, et ils n'auraient aucun intérêt à la défendre, parce que la plupart des résolutions seraient contre eux. La part qu'ils ont à la législation doit donc être proportionnée aux autres avantages qu'ils ont dans l'État: ce qui arrivera s'ils forment un corps qui ait droit d'arrêter les entreprises du peuple, comme le peuple a droit d'arrêter les leurs.

Ainsi, la puissance législative sera confiée, et au corps des nobles, et au corps qui sera choisi pour représenter le peuple, qui auront chacun leurs assemblées et leurs délibérations à part, et des vues et des intérêts séparés."

33. As to this classical form of government, I presented my interpretation in Pasquino (1996, 2009a).
34. On the rationale of the independent exercise of the judicial function, see Pasquino (2001a). The bottom line of the argument seems to be the following: Montesquieu speaking of the separation of powers in England was defending the idea that the agencies that have to exercise the function of applying the laws need to be independent from the agency which exercises the function of making law. Why? To avoid judicial decisions *ad personam*. A *loi* for Montesquieu, likewise for Locke, is/has to be a general abstract commandment—cannot be a *bill of attainder* meaning a norm targeting a specific subject. So the judge cannot make special decisions, since he has to enforce the law that is general and equal for everyone (how is that compatible with a *society of ranks* cannot be discussed here). In this sense the citizen is protected vis-à-vis the extemporary decrees of a biased judge (and moreover he can appeal, at least in the contemporary judicial systems) against a judge's decision which seems arbitrary). The law has to be abstract and general, and the judge independent (tenured) to be able to resist the power of the other branches (for instance, the King), which could try to force the judge to decide in a way that pleases the King. In the case of the CC, the point is different, and I need to be clear about that: the CC is a co-legislator and if the CC is not legally and de facto independent from the political (elected) branches, the CC cannot be a counter-power and its function would evaporate.
35. P. 17; "Il ne suffit pas qu'il y ait, dans une monarchie, des rangs intermédiaires; il faut encore un dépôt de lois. Ce dépôt ne peut être que dans les corps politiques, qui annoncent les lois lorsqu'elles sont faites et les rappellent lorsqu'on les oublie." Montesquieu was referring at the practice of *enregistrement des ordonnances royales* and to the *remontrances* of the *Parlements d'ancien régime*. See Flammermont (1898).
36. I'm avoiding the ambiguous word *power*, but if we understand the exercise of a function as an ability of doing, a power that can be entrusted to different organs or agencies, it is possible to speak of legislative power as the equivalent of the exercise of this paramount function. What matters is to take seriously the split of the legislative sovereign function/power among three independent branches—the point that Madison derived from the "celebrated" Montesquieu and that he adapted to his "republican" (elective) government with two Houses and the President exercising legislative veto.
37. On the classical doctrine of mixed constitution, see the very important books by Nippel (1980) and Blythe (1992).

38. «Séparation des pouvoirs», *Dictionnaire électronique Montesquieu* (2013): http://dictionnaire-montesquieu.ens-lyon.fr/index.php?id=286
39. This idea is already in the text by Kelsen of 1928, and repeated in his answer to Schmitt (Kelsen 2008).
40. I introduced this idea speaking of constitutional courts in democratic societies in a couple of papers some years ago, see Pasquino (1998c, 2006b).
41. See Chantraine (1970: 273):

> **δῆμος** : m., dor., etc. δᾶμος ; d'abord « pays, territoire », cf. *Il.* 5,710 : Βοιωτοὶ μάλα πίονα δῆμον ἔχοντες ; les habitants de ce territoire, cf. *Il.* 3,50 ; déjà chez Hom. (p.-ê. parce que les gens du peuple vivent à la campagne et les grands à la ville), les gens du peuple ; par opposition aux εὐδαίμονες, aux δυνατοί en ion.-att. ; dans un sens politique, en ion.-att. : le peuple souverain, la démocratie, le parti démocratique opposé à ὀλιγαρχία, cf. aussi

42. The reform of the French constitution that, starting from 2010, introduced a mechanism of constitutional adjudication of enacted statutes is probably the most important sign of the expansion of the constitutional democracy (Pasquino 2009b).
43. To be more precise, the Aristotelian mixed *politeia* was a form of government combining elements of two bad forms: oligarchy and democracy. Aristocracy was a good but ideal form, of limited interest for him because of its ideal character.
44. Alito was born to Italian-American immigrants.
45. Sotomayor was born in The Bronx, New York City, and is of Puerto Rican descent. Her father, who had a third-grade education, did not speak English, died when she was nine, and she was subsequently raised by her mother a telephone operator and then a practical nurse (Wikipedia). Antonin Scalia is the son of a Sicilian immigrant. Clarence Thomas was the second of three children born to M.C. Thomas, a farm worker, and Leola Williams, a domestic worker. They were descendants of American slaves, and the family spoke Gullah as a first language (Wikipedia).
46. If one uses this type of rhetoric, it is difficult to understand why the US members of the Senate would be less aristocratic than the judges who exercise judicial review. It is more persuasive and less rhetorical to speak as I do of non-elected, non-accountable public officials.
47. This was the Marxist preoccupation, following the radical model of the French Revolution according to which the Third Estate was all, and Marx

was at the same time against the capitalist oligarchs and against the mixed government somehow revived by the social-democrats.

48. Qualifying as Schmittians the authors who criticize constitutional adjudication is not an easy rhetorical trick to disqualify them; this effortless and commonly used strategy is not what I need in my justificatory theory to show that they are wrong. It is just "to give to Caesar what is Caesar's." I know C. Schmitt's theoretical work well enough to consider disqualifying the reference to his theses and arguments. Even though his answers were often wrong, like the presidential role in the constitution, which American colleagues like Vermuele and Posner (2010) seem to appreciate, his questions are in many cases still with us.
49. See, for instance, Sutter (1997: 139), but already Shapiro (1981).
50. The case of conflict among the central state organs. (See the decision of the It.CC concerning the conflict between the justice minister Mancuso and the Parliament: http://www.giurcost.org/decisioni/1996/0007s-96.htm)
51. English translation on the website of the French Constitutional Council.
52. Starting famously Hume (1748).
53. Hobbes (1651: Ch. XXX): "The office of the sovereign, be it a monarch or an assembly, consisteth in the end for which he was trusted with the sovereign power, namely the procuration of the safety of the people, to which he is obliged by the law of nature [...] But by safety here is not meant a bare preservation, but also all other contentments of life, which every man by lawful industry, without danger or hurt to the Commonwealth, shall acquire to himself.."
54. I presented a systematic interpretation of Hobbes' political theory in three articles (Pasquino 1994b, 2000, 2000b).
55. On Hobbes' method, the following passage from the Preface to the 1647 edition of the *De cive* is relevant: "Concerning my method, I thought it not sufficient to use a plain and evident style in what I have to deliver, except I took my beginning from the very matter of civil government, and thence proceeded to its generation, and form, and the first beginning of justice; for everything is best understood by its constitutive causes. For as in a watch, or some such small engine, the matter, figure, and motion of the wheels cannot well be known, except it be taken in sunder, and viewed in parts; so to make a more curious search into the rights of states, and duties of subjects, it is necessary (I say not to take them in sunder, but yet that) they be so considered, as if they were dissolved."
56. I do not need to suppose any *natural rights*, just those established in a liberal–democratic constitution like the American or the German ones.
57. I do not know of anyone who claims that the representative majority can do whatever it wants (which means a clear rejection of the axiom called

"neutrality" of May's theorem justifying majority rule!). May (1952). The supporters of majoritarian democracy have to suppose that periodical elections and perhaps bicameralism are sufficient mechanisms to guarantee the protection of fundamental rights, though both weak protection mechanisms. Bicameralism with absolute veto power of the two houses in a political system dominated by one political party is of some ambiguous help only if, like in presidential systems, *divided government* is possible. I speak of ambiguous help since this type of system can produce serious gridlock. (The English classical bicameralism with a House of Lords is an effective anti-despotic device, but presupposes a pre-modern, non-equalitarian society.) Elections are certainly an instrument of moderation of the power of the majority of representatives, but only for those who can produce an alternative majority; they can be called pivotal voters.

58. "The Fourth Amendment [of the Hungarian Constitution], adopted 11 March 2013, prohibits the Constitutional Court from examining the substantive constitutionality of future proposed amendments to the Constitution and strips the Court of the right to refer in its rulings to legal decisions made prior to January 2012, when the new constitution came into effect" (from the website of the International Bar Association's Human Rights Institute).
59. "*Vous avez juridiquement tort car vous êtes politiquement minoritaires.*" This sentence was addressed to the conservatives in the National Legislative Assembly by the socialist MP André Laigniel in 1981 at the time of the debates concerning the nationalizations.
60. On majority and truth, see Pasquino (2010).
61. New York Senator Charles Schumer once said during the hearings for a nominee at the USSC: "You will decide about our life and death" (abortion and euthanasia, and now we could add marriage).
62. "If it be said that the legislative body are themselves the constitutional judges of their own powers, and that the construction they put upon them is conclusive upon the other departments, it may be answered, that this cannot be the natural presumption, where it is not to be collected from any particular provisions in the Constitution. It is not otherwise to be supposed, that the Constitution could intend to enable the representatives of the people to substitute their will to that of their constituents. It is far more rational to suppose, that the courts were designed to be an *intermediate body* between the people and the legislature, in order, among other things, to keep the latter within the limits assigned to their authority." Montesquieu attributed a crucial role to the *corps intermédiaires* in his theory of limited/moderated government (see Mosher 2001: 183).
63. One can think of the recent works by Kramer (2004), Tushnet (1999) and Waldron (1999), plus a variety of more recent articles.

64. Worcester v. Georgia, 31 U.S. (6 Pet.) 515 (1832).
65. Sentenza n. 826/1988 and sentenza 466/2002.
66. BVerfG 1995b: 2477. After that decision though: http://web.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=48c44ccb-760c-4869-8e99-fe666bcb4a47%40sessionmgr14&vid=2&hid=24

German High Court Approves New Crucifix Law

Germany's highest court has refused to hear an appeal of a Bavarian court's decision allowing crucifixes to hang in public school classrooms.

The action by the Federal Constitutional Court is the latest in a long-running dispute over the presence of crucifixes in German public schools. In 1995 the German high court ruled that a Bavarian law mandating display of the crucifix in classrooms was unconstitutional, sparking massive protests in the largely Catholic state. Bavarian lawmakers subsequently passed a new law requiring the display of crucifixes but permitting an exception if a parent raises a "serious and reasonable" objection.

67. In any event, the parliament can overrule a decision of the constitutional court amending the constitution. An interesting example is the amendment of the Italian constitution overruling a sentence of the It. CC concerning rules of criminal procedure: the Sentenza 361/1998 canceling art. 513 of the criminal code modified by a statute of August 7th 1997, n. 267, was indeed overridden by the amendment of art. 111 of the Italian constitution passed on November 23rd 1999.
68. For examples in contexts other than US constitutional history, see J. Ferejohn & P. Pasquino, *The Countermajoritarian Opportunity*, 13 (2010) U. PA. J. CONST. L. 353–395.

References

Arndt, Adolf. 1976. *Politische Reden und Schriften*. Berlin/Bad Godesberg: Dietz.

Barthélemy, Joseph. 1906. *Le rôle du pouvoir exécutif dans les républiques modernes*. Paris: Giard et Brière.

Blythe, James. 1992. *Ideal Government and the Mixed Constitution in the Middle Ages*. Princeton: Princeton University Press.

Cappelletti, Mario. 1984. *Giudici Legislatori?* Milano: Giuffrè.

Chantraine, Pierre. 1970. *Dictionnaire étymologique de la langue grecque. Histoire des mots*. Paris: Klincksieck.

Cheli, Enzo. 1961. *Atto politico e funzione d'indirizzo politico*. Milano: Giuffrè.

Cohen, M., and P. Pasquino. 2013. *La motivation des décisions de justice. Le cas des cours souveraines et constitutionnelles*. Paris: Rapport pour le Ministère de la Justice.

Dahl, Robert. 1989. *Democracy and Its Critics*. New Haven: Yale University Press.

Eisenmann, Charles. 2002. *Essais de théorie du droit, de droit constitutionnel et d'idées politiques*. Paris: LGDJ. [notably: « L'esprit des Lois et la séparation des pouvoirs », originally published in *Mélanges Carré de Malberg*. Paris: Libr. du Recueil Sirey.].

Ferejohn, John, and Pasquale Pasquino. 2010. The Countermajoritarian Opportunity. *University of Pennsylvania Journal of Constitutional Law* 13: 353–395.

Flammermont, Jules. 1898. *Remontrances du Parlement de Paris au XVIII^e siècle*. Paris: Imprimerie Nationale.

Gardbaum, Stephen. 2013. *The Commonwealth Model of Constitutionalism. Theory and Practice*. Cambridge: Cambridge University Press.

Ginsburg, Thomas. 2003. *Judicial Review in New Democracies*. Cambridge: Cambridge University Press.

Goldoni, Marco. 2009. *La dottrina costituzionale di Sieyès*. Firenze: Firenze University Press.

———. 2012. At the Origins of Constitutional Review: Sieyès' Constitutional Jury and the Taming of Constituent Power. *Oxford Journal of Legal Studies* 32 (2): 211–234.

Hobbes, Thomas. 1651. *Leviathan*.

Hume, David. 1748. *Of the Original Contract*.

Kaufmann, Erich. 1909. *Auswärtige Gewalt und Kolonialgewalt in der Vereinigten Staaten von Amerika: Eine rechtsvergleichende Studie* ü*ber die Grundlagen des amerikanischen und deutschen Verfassungsrecht*. Leipzig: Duncker and Humblot.

Kelsen, Hans. 1928. La garantie juridictionnelle de la constitution. *Revue de droit public* 44: 197.

———. 1945. *General Theory of Law and State*. Cambridge: Harvard University Press.

———. 2008 [1931]. *Wer soll der H*ü*ter der Verfassung sein?* Berlin: Mohr Siebeck.

———. 2013. *The Essence and Value of Democracy*. Lanham: Rowman & Littlefield.

Kramer, Larry. 2004. *The People Themselves: Popular Constitutionalism and Judicial Review*. New York: Oxford University Press.

Landi, Lando. 1981. *L'Inghilterra e il pensiero politico di Montesquieu*. Padova: CEDAM.

Le Divellec, Armel. 2007. Le gardien de la constitution de Carl Schmitt. In *La controverse sur « le gardien de la Constitution » et la justice constitutionnelle. Kelsen contre Schmitt – Der Weimarer Streit um den H*ü*ter der Verfassung und die Verfassungsgerichtsbarkeit, Kelsen gegen Schmitt*, ed. P. Pasquino and Olivier Beaud, 33–78. Paris: Editions Panthéon Assas.

Lloyd, Sharon. 2014. Hobbes's Moral and Political Philosophy. *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/hobbes-moral/

Machiavelli, Niccolo. 1989. *The Chief Works and Others*. Vol. 1, 101–115. Durham: Duke University Press.

Manin, Bernard. 1989. Montesquieu. In *A Critical Dictionary of the French Revolution*, ed. F. Furet and M. Ozouf. Cambridge: Harvard University Press.

May, Kenneth O. 1952. A set of independent necessary and sufficient conditions for simple majority decisions. *Econometrica* 20: 680–684.

Mosher, M.A. 2001. Monarchy's Paradox: Honor in Face of Sovereignty. In *Montesquieu's Science of Politics. Essays on The Spirit of Laws*. New York: Rowman & Littlefield.

Necker, Jacques. 1792. *Du pouvoir exécutif dans les grands États*, 2 vols.

Nippel, Wilfried. 1980. *Mischverfassungstheorie und Verfassungsrealität in Antike und früher Neuzeit*. Stuttgart: Klett-Cotta.

Parfit, Derek. 1986. *Reasons and Persons*. Oxford: Oxford University Press.

Pasquino, Pasquale. 1994a. Gardien de la constitution ou justice constitutionnelle? C. Schmitt et H. Kelsen. In *1789 et l'invention de la constitution*, ed. M. Troper and L. Jaume, 141–152. Paris: Bruylant – L.G.D.J.

———. 1994b. Thomas Hobbes. La condition naturelle de l'humanité. *Revue Française de Science Politique* 44: 294–307.

———. 1996. Political Theory, Order and Threat. In *Nomos XXXVIII: Political Order*, 19–40. New York: NYU Press.

———. 1998a. *Sieyes et l'invention de la constitution en France*. Paris: Odile Jacob.

———. 1998b. Locke on King's Prerogative. *Political Theory* 26 (2): 198–208.

———. 1998c. Constitutional Adjudication and Democracy. Comparative Perspectives: USA, France, Italy. *Ratio Juris* 11: 38–50.

———. 2000. Th. Hobbes: la condition légale dans le Commonwealth. *Cahiers de Philosophie de l'Université de Caen* 34: 147–164.

———. 2001a. One and Three: Separation of Powers and the Independence of the Judiciary in the Italian Constitution. In *Constitutional Culture and Democratic Rule*, ed. J. Ferejohn, J. Rakove, and J. Riley, 205–222. Cambridge: Cambridge University Press.

———. 2001b. Hobbes, Religion and Rational Choice. *Pacific Philosophical Quarterly* 82: 406–419.

———. 2006a. L'origine du contrôle de constitutionnalité en Italie: Les débats de l'Assemblée constituante (1946–47). *Rivista trimestrale di diritto pubblico* 1: 1–11.

———. 2006b. Voter et juger: La démocratie et les droits. In *L'architecture du droit. Mélanges en l'honneur de Michel Troper*, 775–787. Paris: Economica.

———. 2009a. Machiavelli and Aristotle: The Anatomies of the City. *History of European Ideas* 35: 397–407.

———. 2009b. The New Constitutional Adjudication in France. The Reform of the Referral to the French Constitutional Council in Light of the Italian Model. *Indian Journal of Constitutional Law* 3: 105–117.
———. 2010. Samuel Pufendorf: Majority Rule (Logic, Justification and Limits) and Forms of Government. *Social Science Information*, (Collective Decision Making Rules) 49: 99–109.
———. 2012. Classifying Constitutions: Preliminary Conceptual Analysis. *Cardozo Law Review* 34: 999–1019.
———. 2013a. Democracy: Ancient and Modern, Good and Bad. In *Democracy in a Russian Mirror*, ed. A. Przeworski, 99–118. Cambridge: Cambridge University Press.
———. 2013b. Majority Rules in Constitutional Democracies. Some Remarks About Theory and Practice. In *Majority Decisions*, ed. Stéphanie Novak and Jon Elster, 134–156. Cambridge: Cambridge University Press.
———. 2015. Disclosed and Undisclosed Voting in Constitutional/Supreme Courts. In *Secrecy and Publicity in Votes and Debates*, ed. Jon Elster. Cambridge: Cambridge University Press.
Przeworski, Adam. 1999. Minimalist Conception of Democracy. A Defense. In *Democracy's Value*, ed. I. Shapiro and C. Hacker-Cordon, 23–55. Cambridge: Cambridge University Press.
———. 2011. *Democracy and the Limits of Self-Government*. New York: Cambridge University Press.
Sartori, Giovanni. 1987. *The Theory of Democracy Revisited*. Chatham: Chatham House.
Schmitt, Carl. 1931. *Der Hüter der Verfassung*. Berlin: Dunker and Humblot.
Shapiro, Martin. 1981. *Courts: A Comparative and Political Analysis*. Chicago: University of Chicago Press.
Simon, H. 1994. Verfassungsgerichtsbarkeit. In *Handbuch des Verfassungsrechts*, ed. E. Benda, H. Maierhofer, and W. Vogel, 1637–1681. Berlin: de Gruyter.
Smend, Rudolf. 1923. *Die politische Gewalt in Verfassungsstaat und das Problem der Staatform*. Tübingen: J.C.B. Mohr (P. Siebeck).
Strauss, Leo. 1995 [1932]. *Notes on Carl Schmitt, The Concept of Political*. Trans J. Harvey Lomax. Chicago: University of Chicago Press.
Strawson, P.F. 1959. *Individuals – An Essay in Descriptive Metaphysics*. London: Methuen.
Sutter, Daniel. 1997. Enforcing Constitutional Constraints. *Constitutional Political Economy* 8: 139.
The New Encyclopedia Britannica. 1993. *Micropaedia*, 15th ed. Vol. 4. Macmillan: New York.
Tushnet, Mark. 1999. *Taking the Constitution Away from the Courts*. Princeton: Princeton University Press.

Vermuele, Adrian. 2011. Second Opinions and Institutional Design. *Virginia Law Review* 97: 1435.
Vermuele, Adrian, and Eric Posner. 2010. *The Executive Unbound: After the Madisonian Republic*. Oxford: Oxford University Press.
Vinx, Lars. 2015. *The Guardian of the Constitution. Hans Kelsen and Carl Schmitt on the Limits of Constitutional Law*. Cambridge: Cambridge University Press.
Von Bruneck, A. 1988. Constitutional Review and Legislation in Western Democracies. In *Constitutional Review and Legislation: An International Comparison*, ed. C. Landfried. Baden-Baden: Nomos.
Waldron, Jeremy. 1999. *Law and Disagreement*. Oxford: Oxford University Press.
Weiler, J.H.H. 2011. In *The Worlds of European Constitutionalism*, ed. G. de Burca and J.H.H. Weiler. Cambridge: Cambridge University Press.
Zakaria, Fareed. 2003. *The Future of Freedom: Illiberal Democracy at Home and Abroad*. New York: W.W. Norton & Co.

Assessing Constitutional Efficacy: Lessons from Mexico's Hegemonic Party Era

Andrea Pozas-Loyo

Introduction

Mexico was governed by a hegemonic party system centered in a powerful executive from 1929 to 2000 when it lost the presidential election. During these 70 years, the PRI (*Partido Revolucionario Institucional*) had control over the administration, the Federal Congress, the state governments, and the judiciary. The president was the cornerstone of a well-disciplined political system: he was the head of the government and the head of the party. The president had the political capacity to go around some provisions of the 1917 Constitution without political opposition,[1] or rather to amend much of it due to the PRI's supermajoritarian legislative control.[2] Nevertheless, this does not imply that in this period the president could transform in an

For Russell Hardin, in memoriam.
His generosity and integrity both academic and personal have been, and will continue to be, a motivational force.

A. Pozas-Loyo (✉)
Instituto de Investigaciones Jurídicas, Universidad Nacional Autónoma de México, México, DF, Mexico

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_9

arbitrary fashion *any* constitutional article. In particular, during this president-centered era, Article 83 of the constitution that establishes a 6-year presidential term without reelection was neither altered nor violated, by any president. Without doubt this constituted a very strong constraint on power on otherwise powerful individuals. Why presidents could not change Article 83 nor violate it? Was Article 83 efficacious? How can we know? And, what lesson can we draw from this case about how to assess constitutional efficacy in general?

To answer these questions, I analyze President Miguel Alemán's (1946–1952) unsuccessful attempt to seek reelection. In particular, my account focuses on answering, "Why President Alemán failed?" Why was not he able to change or go around Article 83 to reelect himself or extend his tenure as he and many other important politicians at that time so wished? I argue that the mechanisms that protect a constitutional norm from ad hoc change or violation and their relation to the constitution are the key to assess the norm's efficacy. I show that understanding those mechanisms enables us to evaluate whether and to what extent Article 83 was efficacious in Mexico at the end of President Alemán's tenure.

The reminder of the chapter has four sections. In the first one, I present and discuss President Alemán's unsuccessful attempt to change the constitution to be reelected for a second term. My account focuses on the determinants of his failure, that is on the mechanisms of constitutional enforcement at play. In the second section, I discuss whether and to what extent Article 83 was efficacious and reflect on constitutional efficacy in autocracies and on how to assess it more generally. I claim that to determine the degree of constitutional efficacy of a constitutional norm, it is necessary to identify the mechanisms of constitutional enforcement and their relation to the constitution. An as a generalizable test, I distinguish three levels of constitutional efficacy based on the relation between the mechanisms of constitutional enforcement and the constitution: (1) cases of parallel enforcement mechanisms, where there is mere text-reality coincidence; (2) cases of exogenous constitutional enforcement, where the constitution plays an important role (i.e. functions) but it is not strictly efficacious; and (3) cases of endogenous constitutional enforcement, where there is strict constitutional efficacy. In the third section, I briefly conclude.

Enforcing Alternation of Power Under Autocracy: Alemán's Failed Reelection Bid

Mexican Constitution of 1917, Article 83: [...] The citizen who had performed as President of the Republic, popularly elected or under the interim or alternate character, or provisionally takes the office of the Federal Executive, in no case and under no circumstances may perform again this position.

On the Principle of No Reelection in Mexico

After 7 years of a bloody war, the 1917 Constitution became the best political answer to return to the constitutional order, achieve peace, and consolidate the constitutionalist coalition victory in a country where there were still many uncertainties (Marván 2017). Whereas in the constitution-making process there were some disagreements (Marván 2017, 122–152), the principle of no executive reelection was unquestionable. To understand the significance of this principle in Mexico's history, some context is necessary. Since the nineteenth century, there was an extended belief that presidential reelection led to dictatorship. Hence, its prohibition was present in the *Plan de Tuxtepec (1876)* that had been Porfirio Díaz's ideological manifest; however, once president he betrayed his commitment and was reelected for nine periods, 30 years. Against this background, and with the motto "effective suffrage and no reelection," the Mexican Revolution started in 1910, forcing Díaz into exile. Thus, the unchallenged absolute prohibition of executive reelection in the 1917 Mexican Constitution: it was considered a necessary institution to leave behind the great evils of strong men's dictatorship and political instability.

After a failed attempt in 1925, in 1927 and 1928, General Álvaro Obregón (president from 1920 to 1924) was able to amend Article 83 to enable non-consecutive reelection allowing him to run for a second term, but as would be stressed when President Alemán attempted reelection, Obregón was assassinated before taking office. In 1932, the National Convention of the National Revolutionary Party, predecessor of the PRI, proposed to reform Article 83 to its original form, with an absolute and clear prohibition of presidential reelection. Since then it has never been amended again (Barquín Alvarez 1990, 195–198). Nevertheless, from 1928 to 1934, Article 83 was hardly efficacious, since President Plutarco Elias Calles had de

facto kept the executive power while placing political puppets in the presidency.

Everything changed in 1934, when General Lázaro Cárdenas was elected president. President Calles had chosen General Cárdenas for the presidency, believing that he, as his predecessors, would be easily manipulated. President Cárdenas turned up to be an extremely capable politician; he acquired independence political force, eventually forcing Calles to exile in the United States in 1936. Lázaro Cárdenas was the first president to de facto hand in the presidential power at the end of his term; after him the absolute prohibition of executive reelection has never been violated nor changed.

President Cárdenas consolidated a political system that enabled the hegemony of a single party until 2000. In particular, even though he had amassed great political power, respect, and popularity, when his tenure came to an end, he did not seek reelection and instituted the intra-party mechanism by which Mexican presidents would be de facto selected until 1994: the serving president's last function was to choose the next presidential candidate (knowing that he would be elected) after consultation with the ideologically very diverse power groups that formed the party (Cosío Villegas 1975). So in 1940 General Manuel Ávila Camacho was elected president. President Cárdenas had picked him to be the official candidate, over his closest allies, who would have continued his policies. Cárdenas arguably sensed that a more center-right politician would enable political stability given the inconformity his leftist policies had caused among powerful social and political sectors. In this way starting with President Cárdenas, there was a periodic rotation of presidential power among the very different groups that constituted the hegemonic party. As Casanova explained in his 1985 account of the hegemonic party system, "Once the selection of the presidential candidate has been resolved, the new composition of forces produces a renovation of the directive positions, generally permitting a more realist reflection of the nature of the power coalition" (Cassanova 1985, 70).

Miguel Alemán was the first civilian to become president after the Revolution, in 1946. During his tenure "authoritarianism was modernized" (Medina 1982). He strengthened the mechanisms of power control within the party,[3] over Congress, state governments, and socially active groups such as unions and organized peasants, using force at times (Medina 1982; Torres 1984; Servín 2002). For this reason, and for his support to the United States during the Cold War, he was resented by the left wing of the party, linked to President Cárdenas. In words of Lombardo

Toledano, a very influential union leader who, as we will see, strongly opposed Alemán's reelection:

> During Alemán's government the workers did not dare to do large strikes because Alemán was completely on board with the cold war. . .it was the worst period we have went through. . .President Alemán is historically responsible for intervening for the first time in internal government of unions. It was he who ordered the military occupation of the railway union, deposed the organization's executive committee, and later did the same with the great industrial unions. (Wilkie and Monzón 1969, 314)

In the last months of 1949, after the death of Gabriel Ramos Millán who many thought would become the PRI presidential candidate, some of the closest allies of outgoing President Alemán launched an open campaign to promote his reelection. In what follows, I present a brief chronologic account of the months that this campaign lasted, from November 1949 to October 14, 1951. My aim is to answer why the attempt to change the 1917 Constitution to enable Alemán's reelection failed? Through which mechanisms was the constitution enforced? Can we claim Article 83 efficacious, and why?

The Reelection Campaign

The first point I want to defend is that the attempt to amend the constitution to enable executive reelection can be ascribed to President Alemán. In other words, that it cannot be considered a campaign organized entirely by his subordinates without the president's support, as he and other politicians claimed at the time, and later (e.g. Alemán 1987, 242, quoted in Chanes Nieto 1993, 154). This point is important for my account on Article 83's enforcement not only because it takes as a premise Alemán's support to such a constitutional change but also because the mere fact that President Alemán did not explicitly accepted that he backed this very public campaign (encouraged by his behavior) already tell us that the constitution matter in some important way.

As Przeworski puts it, at a minimum the law can matter as does a pole streetlight: it may not alter the destination you want to reach, but you need to at least circumvent it (personal communication). This was the case of Alemán's public discourse on reelection early in this campaign, at a time he did not know whether it would succeed: Alemán publicly claimed respect for the prohibition on reelection while his and his

subordinates' behavior aimed at circumventing it. If Article 83 of the constitution would not have mattered, if it would not have had any normative weight at all, Alemán would have publicly recognized and supported his amendment project, as he did with the other 20 constitutional amendments his administration successfully passed (Fix-Fierro and Valadés 2016). This was not the case with regard to Article 83. In December 1949 he declared: "I have never thought of the possibility of a reelection in my favor. I do not encourage actions that are in disagreement with the legal principles that govern us" (Chanes 1993, 155).

As several historians (e.g. Cosío Villegas 1975; Chanes 1993; Sevin 2002) and politicians of that period (e.g. Wilkie and Monzón 1969) have argued, claiming that Alemán had nothing to do with the attempt to change Article 83 of the constitution is unfeasible: in this period no executive official would have dared to take such an important decision and back it with such a very well-organized campaign without the presidential support. In addition, the two leading figures of the campaign were high-ranking officials very close to the president: Santiago Piña Sorio, the Director of the Joint Chief of Staff (el jefe del Estado Presidencial), and Rogerio de la Selva, the Private Secretary of the Presidency (Secretario Particular de la Presidencia) (Cárdenas 1973, 440; Wilkie and Monzón 1969, 365; Servín 2001, 120; Medina 1982, 163). Moreover, Alemán's political behavior and how it was interpreted at the time backs the hypothesis that he was behind this campaign.

The press played a central role in the PRI Hegemonic Era: it was the space where many of the political messages among different political groups were sent. As I have already mentioned, the left wing of the party, linked to President Cárdenas, opposed President Alemán's policies and this opposition was crucial for the enforcement of Article 83. By the end of 1949 and the beginning of 1950, the campaign to amend the constitution was well known, and not surprisingly, the tensions between Cardenists and Alemanists were being displayed in the national press.

In an extremely uncommon public display of his political opinions, former President Cárdenas published a declaration against reelection in a national newspaper. In his diary, he selected the following paragraph from such publication:

> Lessons from History oblige us to maintain the antireelectionist tradition, the vitality of our people requires the renovation of his men over any rule by 'a strong men' (*caudillaje*) that is so detrimental to democratic effectiveness and

> to the progress of the nation. In this way, I consider myself respectful of these traditions that nurture the civic life of our own people. (Cárdenas 1973, 378–9)

This publication was immediately followed by a series of newspaper reports with harsh criticisms of Cardenas' administration characterizing it as dishonest and irresponsible (e.g. Mendieta Núnez 1949). These publications were perceived as coming from Rogerio de la Selva, the Private Secretary of the Presidency (Servín 2002, 133). The tension grew after the PRI's National Assembly in February 1950, in which several institutional changes were enacted to strengthen party discipline in face of the coming presidential succession such as the preclusion of local civic committees for the discussion of party candidates (Servín 2002, 129).

However, the signals of party union and discipline did not stop the publicity of the confrontation. On April 15, 1950, 25 former Cárdenas' collaborators published a whole page in *El Univer*sal, a leading national newspaper stressing that President Cárdenas "had strictly followed the principles of the Revolution" (i.e. he had not reelected himself) and that being a Cardenist meant "the identification with a presidential term...characterized by the loyal compliance with the constitutional commands that give structure to the nation..." (Angulo et al. 1950). It is important to note that several of the signers were still very important figures in the armed forces and the public sphere; therefore, such a publication carried an important political weight. Cárdenas noted in his diary on April 17th that he had been told that General Jara had organized the publication, and that former President Ávila Camacho (1940–1946) was informed of it before its publication. As we will see, former presidents played an active role as enforcers of Article 83.

Five days later, on April 20 several prominent politicians from the left visited the president of the PRI (General Taboada)[4] "to express their conviction that the 'revolutionary left' could only be part of the PRI, that the rumors that they wanted to create a new political organization were unsubstantiated." In such a disciplined party, just acknowledging "rumors" of a possible split was highly unconventional. Interestingly, as Servín notes, Charles Burrows, analyst of the US Department of State, interpreted the visit to the PRI president in the aftermath of the publication of the "Cardenist manifest" as a strategy to stop reelection "in which former President Cardenas may be involved" (Burrows NAW, RG59, 712 00/4-1850, quoted in Servín 2002, 133).

By June 1950 the reelection campaign acquired more intensity. In April Congressman Rafael Ortega, General Secretary of the Mexican Confederacy of Workers and Peasants (*Confederación Obrera y Campesina de México*), proposed in the assembly of that organization to explicitly support President Alemán's reelection, and the proposal was approved by acclamation, "in the following months groups with similar proposals appeared in different parts of the country" (Alemán 1987, 386). On June 3, the National Confederation of the Family made a formal request to the Federal Congress to amend several provisions of the constitution in order to enable the reelection of President Alemán (Hoy June 3rd 1950; Servín 2002, 121).

During the summer of 1950, two political parties were formed to support President Alemán's reelection. In June the formation of the Political Party "Constitutional Article 39" was announced. Article 39 of the 1917 Constitution states: "The national sovereignty is vested, originally and essentially, in the people. Public power comes from the people and it is institutionalized for the people's benefit. The people, at all times have the inalienable right to change or modify its form of government." The leaders of the new party claimed that this Article constitutionally grounded the possibility of reelection if backed by a popular mandate. Among the leaders of this party were Guillermo Ostos who was part of Alemán's cabinet (Medin 1990, 163). Additionally, in July the National Reelectionist organization tried to register as a party; according to a report of a British foreign office analyst, the leaders of the organization claimed to have 45,000 members including 3 generals, 2 senators, and 26 congressmen among other public servants (Fisher, PRO, FO371 quoted in Servín 2002, 122).

Moreover, in 1950 several Congressmen such as Alfonso Reyes Hernández explicitly backed Alemán's reelection. In this context, some governors opportunistically also sent a signal of support. This was the case of the governor of the state of Morelos who passed a law extending the presidential term in his state for 2 years. This law was declared unconstitutional by the Supreme Court, of course, since it was intended merely as a signal of political support not a legal norm that could be taken seriously (Cosío Villegas 1975, 119–120). The president of the PRI, General Taboada, threatened to expel from the party whoever supported those reelectionist parties, but the public display of political strength backing constitutional amendments in favor of presidential reelection was already made.

In May 1950, President Alemán's public exposure remarkably increased. He toured the south of the country, a series of performances that were

described by an analyst of the US State Department and by the British Foreign Office as a "candidate's tour during an electoral campaign" (Burrous, NAW, RG59, 712.00/6-650; Fisher PRO, FO371 quoted in Servín 2002, 122). It is noteworthy that campaign tours were very important during the PRI Hegemonic Era. Of course they did not aim to gain the popular vote in a competitive election, but they were crucial for the PRI's candidate since through them he consolidated links and created two-way commitments with the states' political elite (Pozas-Horcasitas 2009). That tour was followed by another presidential tour in the north of the country. Both tours included a large number of cabinet members and other important political figures. In Monterrey, the second most important city in the country, the walls, and principal avenues were covered with publicity in favor of President Alemán's reelection (Servín 2002, 122–24).

Last, but by all means not least, the campaign in favor of Alemán's reelection targeted the Armed Forces. On June 11, 1950, Cárdenas wrote in his diary:

> Today Major General Federico Montes with whom I have an old friendship visited me, and told me that he was required by the chief of the Security Services of the Republic's Presidency, Marcelino Inurreta, to sign a declaration of allegiance to President Alemán and a commitment to backup any constitutional reform in favor of re-election or presidential tenure extension. He also told me that he saw the document signed by the Generals Pedro Villaseñor, Lucas González, Aguille Manjarrez, Tomás Sánchez Hernández and others. He added that he and General Alejo Gonzáles refused to sign. (Cárdenas 1973, 399–400)

Montes also told Cárdenas that he and other generals worried for the state of affairs in the country, that they had decided to take an active role the following presidential election, and that they had told so to President Alemán (Cárdenas 1973, 400).

Before focusing on the responses that the previous behaviors elicited, let me briefly note the international context in which the reelection campaign took place, since, as we will see, it had a significant impact on Alemáns' perceived possibilities to reach his aim. On July 3, 1950, the United States had mobilized its troops to Korea. The beginning of the Korean War and the intensification of the Cold War were welcomed by the right in Latin America. Miguel Alemán was known by his affinity to Truman's policies, and in this context, it was expected that if successful his reelection would be backed by the United States. In words of Lombardo Toledano[5]:

> [The Korean War] was the cause of the political turn to the right that Latin America and of the *Coups d' Etat* [of the period]. Even the President of Costa Rica, doctor Calderón Guardia, who was a catholic, was considered a communist and an armed movement was organized against him. (Wilkie and Monzón 1969, 368)

In his long interview with Wilkie and Monzón, Lombardo Toledano describes a discussion he had with President Alemán that captures the importance of the international context on the reelection attempt.

> [When I realized the reelection campaign was for real] I went to talk with Alemán and told him "It is nonsense". "Why?" he asked, "Because your reelection is not possible, the Constitution needs to be amended." "Well, but General Obregón was reelected", he responded. "Those were other historical conditions -I replied- you cannot try it, you will fail, I know why you are attempting to be reelected because President Truman told you and the other Latin-American Presidents that the third war may happen in months...González Videla, President of Chile, declared so to a Brazilian newspaper...." (Wilkie and Monzón 1969, 368)

In this context, the right wing of the PRI felt empowered and the left threatened. There were some editorials asking for action against leftists to fight "Communism" (Cárdenas 1973, 418). There was a growing concern that acts of repression would be "legitimized by the war" (Cárdenas 1973, 417).

The Response to Alemán's Reelection Attempt

The opposition to any constitutional reform that would enable Alemán's reelection was clear and strong and came from a diversity of fronts. It was also very public, and while the language maintained the standards of "political correctness" of the regime,[6] the signals and their political weight were by all understood.

On June 17, 1950, General Sánchez Taboada, the president of the PRI, gave a press conference in which he claimed that President Alemán and the PRI were opposed to reelection and that the constitution would not be amended. "The President would maintain his respect for the revolutionary principle 'effective vote no reelection'" (*Hoy*, June 17).[7] It is hard to overestimate the political significance of this press conference. The very fact that the president of the PRI felt the need to publicly and strongly

oppose any constitutional amendment to permit reelection, a principle that had been taken for granted since Cárdenas, was extraordinary. The cautionary undertone was by no one missed. As we have seen, that summer President Alemán had been touring the country and the reelection campaign was in its higher point gaining impulse by Korean War.

General Taboada was not the only powerful party leader to publicly oppose the possibility of Aleman's reelection. As already discussed at least two generals had refused to sign the letter in support to an eventual constitutional change, and several generals had met with President Alemán to express their concern for the campaign supporting reelection as well as their intent to play an active role in the succession period.

Both President Cárdenas and President Ávila Camacho were also highly respected generals, with many strong ties in the Armed Forces, the unions, and the political elite of the PRI and also of the opposition. To understand the dynamics of the hegemonic party, it is crucial to know that the Mexican political elite was a very dense network, where individuals not only had many ties, but those ties were of different kinds. For instance, during the months that the reelection campaign lasted, President Cárdenas met with President Alemán three times, twice in a dinner with their wives and once in an official event. During these months, Cárdenas also met twice with former President Ávila Camacho and once with former President Ortiz Rubio; he met with several generals, ambassadors, governors, and Congressmen; and crucially, as I will discuss later, he met several times with General Henríquez and with Vicente Lombardo Toledano, both of whom decided to run for president as a response to Alemán's reelection campaign.

Former Presidents Cárdenas and Ávila Camacho met on June 20, when Ávila Camacho stayed overnight at Cárdenas' home in Michoacán. According to Cárdenas' account, they discussed the "reelection issue" and Ávila Camacho expressed his opinion that "despite the reelectionist propaganda that has been undertaken within the official sphere, he considers that President Alemán will [ultimately] reject the insinuations for his reelection..." (Cárdenas 1973, 401). Remember that Cárdenas had "selected" Ávila Camacho as the PRI candidate back in 1940, and that Ávila Camacho had done the same with President Alemán in 1946. Therefore, it is not surprising that Ávila Camacho reassured Cárdenas that Alemán would ultimately respect the no reelection constitutional norm. Nevertheless, as Servín stresses, 2 days later President Ávila Camacho felt the need to make public such "trust," and in a very unusual interview he stated:

> I do not believe that in Mexico there will be a new reelection. I know the feelings of President Miguel Alemán and his antireelectionist convictions, therefore the efforts that his collaborators do in this respect. . .to re-elect him will be useless. Antireelectionism has helped our country in its development, enabling the renovation of men. Antireelectionism must be maintained in Mexico as an example for the whole world, an example that, if followed, would resolve many problems in Latin America. . .Antirrelectionism is one of the great conquests of the Revolution and without doubt one of the main motors of our economic development. (Excélsior, June 22nd quoted in Servín 2002, 127)

Moreover, according to Gustavo Espinosa Mireles, who was present during the interview, Ávila Camacho noted, "the only one who broke this constitutional prohibition, General Álvaro Obregón, was killed for doing so . . ." (Servín 2002, 127). While this strong comment was not published, it was noticed. As already mentioned, Cárdenas himself had also broken months before the informal rule that maintained former presidents out of the public eye, publishing in a National Newspaper his opposition to reelection. The extraordinary public statements were arguably only the tip of the iceberg: the PRI political elite was under turmoil.

The reelection attempt created a deep divide within the PRI and activated the opposition. General Cándido Aguilar decided to split from the PRI and agreed to run as an independent candidate for the presidency. General Cándido Aguilar was a respected military officer with excellent "revolutionary credentials"; he was the son in law of Venustiano Carranza (the leader of the Constitutionalist Army, which won the Revolution and enacted the 1917 Constitution). Cándido Aguilar had been nominated by Alemán to be part of the Legion of Honor of the Mexican Arm Forces and he had important political influence in the states of Veracruz and Tamaulipas.[8]

General Miguel Henríquez Guzmán also decided to split from the PRI, to run for president. He was also a renowned military officer, and his brother was a very successful businessman who put his economic resources behind Henríquez's campaign. Henríquez campaign was able to mobilize a substantial number of people: by June 1950 there were already 22 Henriquista local committees in 10 states and the capital, and by 1951 the Henriquista movement had presence throughout the territory (Servín 2002, 136). Several important figures from the PRI split from the party and supported Henríquez's candidacy. Last but not least, Henríquez was a very

close friend of President Cárdenas. There was a generalized perception that President Cárdenas backed Henríquez campaign, even if Cárdenas never explicitly said so. Henríquez often visited the former president, and several members of the Cárdenas family attended Henríquez's rallies, which filled large squares all over the national territory. To make clear how Cárdenas used the ambiguous relation to Henríquez's campaign, and how such ambiguity worried Alemán, it is noteworthy to quote Cárdenas' notes on a conversation he had with President Ávila Camacho in June 1950:

> His [President Ávila Camacho's] conversation extended letting me know that "in Mexico" it's been said that friends of mine "proclaim" that they work in favor of general Henríquez with my authorization. And that he feels that Mr. Miguel Alemán is no friend of General Henríquez. I thanked his conversation, and manifested that those versions were natural in the political context in which the country is now, that my apolitical attitude stands invariable. That I am a friend of General Henríquez and that he is as well. (Cárdenas 1973, 401)

In addition to Cándido Aguilar and Henríquez, Vicente Lombardo Toledano also decided to run for president. In June 1950 Cárdenas wrote that Lombardo Toledano had paid him a visit and communicated his decision to take part in the elections:

> On the country's politics he spoke about intention of being candidate to the Presidency of the Republic. That he admits he won't win the electoral fight, but [he thinks]...it will serve as platform to enhance the faith [hacer fe] in the principles of the Mexican Revolution. (Cárdenas 1973, 400)

Several years later Lombardo Toledano spoke of his candidacy in the following terms: "I knew very well I would not [win]...but the campaign opened a perspective that would become a reality in President López Mateos' term" [i.e. a new turn to the left] (James Wilkie y Edna Monzón 1969, 374). Finally, the PAN (*Partido Acción Nacional*), the historical opposition party from the right, nominated Efraín González Luna as its presidential candidate.

By June 1951 the possibilities of a constitutional amendment to permit reelection were vanishing. Probably as a last resource, on September 12, 1950, Alemán's Private Secretary sent General Adalberto Tejada and Gonzalo Vázquez Tejeda to speak with former President Cárdenas. Cárdenas described such visit in the following terms:

> . . .General Tejada told me "Excuse us, Mr. Rogerio de la Selva . . .wishes to know which is your opinion on the President's reelection" . . . I made clear . . .that they could make it [my opinion] public. I consider that only false friends of President Alemán wish him to be reelected. I recognize enough intelligence in him not to admit his continuity leading the government, and that he will know how to contribute, with his example, to strengthen the democratic principles. . . that he will not permit that the false theory of indispensable men in power would again be nurtured. . ..*Reelection, in the best of cases leads to dictatorship and dictatorship causes violence. . .Mexico must be guard of new civil wars.* . ..(Cárdenas 1973, 440)

The threat was real. As Cosío Villegas, an influential historian and intellectual of the period, put it:

> One can suppose that Don Miguel (President Alemán) weighed in the resistance to his permanence in power and even the serious risk that Cárdenas and other great personalities would decide to move the opposition to the terrain of arms, and that they would have an excellent flag to make a military movement succeed. There was a real proof of such a danger: The candidacy of General Miguel Henríquez Guzmán started to be supported by recognized Cardenists and even by members of the family of the General [Cárdenas]. (Cosío Villegas 1975, 120)

Two days later, Adolfo Orive visited Cárdenas in the name of President Alemán to inform him that the "official milieu" was leaning in favor of the candidacy of Adolfo Ruiz Cortines and that the continuation of President Alemán leading the government will come only in case of an international conflict that affected Mexico (Cárdenas 1973, 441). A month later, the PRI convention nominated Adolfo Ruiz Cortines as its candidate for the presidency for the period 1952–1958.

To close this section, I want to note that the reelection attempt had as one of its many consequences the nomination of Adolfo Ruiz Cortines as candidate for the presidency. Ruiz Cortines was clearly picked as a conciliatory move; he was considered a moderate, earnest, and austere politician who was not close to Alemán, and therefore could built bridges among the different resentful political groups that still constituted the PRI.

Was Article 83 Efficacious? How Do We Know?

In this section I discuss whether and to what extent was Article 83 efficacious and reflect on what lessons can we draw on constitutional efficacy in autocracies and on how to assess it more generally from the case analyzed in detail in the previous section.

On Constitutional Change and Enforcement in Autocracies

The first inference from our case is that the common claim that nondemocratic regimes' dynamics of constitutional change precludes the possibility of efficacious constitutional constraints in those regimes is false (see also Barros 2002, and Pozas-Loyo and Ríos-Figueroa 2017). According to this claim, in nondemocratic regimes the executive always has the capacity to make ad hoc constitutional amendments (i.e. make at will constitutional changes to serve its interests), and therefore the constitution and its constraints cannot be efficacious vis-à-vis the executive's behavior. In Tushnet's words: "the authoritarian leader has lawful power to alter constitutional provisions at will..." (Tushnet 2015: 425).

This claim is an implication of a familiar conceptualization of "authoritarianism": "I take as a rough definition of authoritarianism that all decisions can potentially be made by a single decision maker [and that] those decisions are [...] unregulated by law" (Tushnet 2015, 448). In other words, by this definition, "if the regime is authoritarian, it faces no constraints on abandoning law, courts, and constitutionalism, when doing so would serve the regime's interests..." (Tushnet 2015, 432). Therefore, this argument excludes *a priori* the possibility of efficacious constitutional constraints on authoritarian executives since by definition they always have the capacity to amend the constitutional provisions at will (Tushnet 2015, 425).

The failure of President Alemán to amend Article 83 of the 1917 Mexican Constitution is, I believe, a counterexample to the above argument. As I showed in the previous section, President Alemán *could not change* the constitution to enable his reelection. And therefore, since ad hoc *constitutional change* was not possible regarding Article 83, we can conclude two things: first, that in this case, we cannot *a priori* preclude the possibility of the efficacy of Article 83. And second that the initial claim, which *a priori* denies possibility of efficacious constitutional constraints of the executive in authoritarian regimes, is not generalizable over all nondemocratic cases.

Assessing Constitutional Efficacy Using Enforcement Mechanisms

But was Article 83 efficacious? How can we know? Of course, the first step to answer these questions is to provide an account of "constitutional efficacy." I have elsewhere discussed this issue at length, claiming that when constitutional roles are invested in an individual, she receives special kinds of motivations, which I call "constituted motivations." The account of constitutional efficacy I defend is understood to be the prevalence of those motivations in the behavior of individuals holding constitutional roles (Pozas-Loyo 2012).[9] Now, of course, there is an observational problem to assess constitutional efficacy, so understood: if it is determined by the kind of motivations that cause constitutional role holders to behave in agreement with constitutional norms, how can we know if a given constitutional norm is efficacious if we can only observe whether behavior is in agreement with the norm, but have no access to what motivated such behavior? In other words, how can we assess constitutional efficacy given that motivations are not observable and behavior consistent with constitutional norms can be motivated by very different factors?

Here I want to argue that through the study of the enforcement mechanisms of constitutional norms and their relation to the constitution, we can approximate the nature of the motivations behind behavior consistent with constitutional mandates. In other words, I claim that to assess the degree of constitutional efficacy of a constitutional norm, we can approximate the motivations by identifying the mechanisms of constitutional enforcement and their relation to the constitution. By "enforcement mechanisms of constitutional norms," I mean the factors that encourage behavior consistent constitutional norms, that is to say, the sources of costs or benefits that when known or believed by a person produce individual motivations, which lead to a behavior in agreement with constitutional prescriptions.

To clarify this point, let me identify the enforcement mechanisms that were at play in Alemán's failure to be reelected. In the account of President Alemán's impossibility to ad hoc amend or violate Article 83, two enforcing mechanisms can be identified: first, those linked to the intra-party opposition to Alemán's reelection led by President Cárdenas, President Ávila Camacho, and General Taboada. Given the political and social capital of these three leaders, particularly their strong connections with the Armed Forces and diverse social and political organizations, their capacity to infringe huge costs over Alemán was considerable, and Alemán knew so. Moreover, they were emphatic and public about their opposition to

reelection, and the threatening undertones sent the message that action could be expected if Article 83 was not respected.

The second kind of enforcement mechanisms came from outside the PRI. In particular, the opposition formed by former members of the party who had decided to split from it were capable of producing high costs over Alemán. As we have seen, Henríquez candidacy was considerably popular, and more importantly it had the possibility to grow a lot if the left wing of the PRI decided to unify behind it in the face of a constitutional violation or amendment to enable presidential reelection. Such move was not unfeasible given that Henríquez was close to Cárdenas, and the latter still was the moral authority of the left. Moreover, the splits the PRI suffered were a vivid reminder that the party's integrity depended on the possibility of power alternation among the different ideological groups. Only as long as the alternation of presidential power within the party was possible (i.e. as long as no president sought reelection), no group would break with the party and all would respect the candidate selection. In sum, these two enforcement mechanisms arguably grounded in President Alemán a justified belief that the costs of pursuing reelection would be too high.

Now, why and how exactly can we approximate the motivations leading to behavior consistent with constitutional norms by identifying the mechanisms of constitutional enforcement and their relation to the constitution? To clarify this point, let me distinguish three levels of constitutional efficacy based on the relation between the mechanisms of constitutional enforcement and the constitution:

1. Cases of parallel enforcement mechanisms: mere text-reality coincidence
2. Cases of exogenous constitutional enforcement: the constitution functions
3. Cases of endogenous constitutional enforcement: constitutional is efficacious

A codified constitution is a system of norms. It is a system because its constitutional provisions are interrelated, creating a more or less consistent whole. And that system is of norms because its provisions establish constitutional roles (e.g. that of Supreme Court Justice or President) and regulate the behavior of individuals occupying those roles. But, codified constitutions are not the only normative systems of political life. Historically, in fact, they are latecomers: they have been present in the political scene only since

the late eighteenth century. Moreover, even in countries with codified constitutions, the constitution is only one among many political normative systems that can potentially regulate interactions of individuals who happen to be in constitutional roles. Furthermore, politics is not an isolated sphere, and normative systems are present in all areas of social life. In this way, a complex net of normative systems constitutes social and political life (Searle 2010).

Now, any given individual has a number of different roles. For instance, an individual with a constitutional role like that of "the President" can also be member of a party, a corporation's stoke holder, a friend of many, and a parent of two. And, therefore, a given interaction between two individuals holding constitutional roles can be regulated by a number of different, potentially conflicting, normative systems (Merton 1949). For instance, an interaction between two individuals holding the constitutional roles of "vice-president" and "member of Congress" correspondingly could be regulated by a constitutional provision linked to those roles, by an informal corporative norm if they both are board members of a corporation, and by an interpersonal norm if they happen to be friends, among many others.

Here I am interested on what I call parallel norms. This is its definition: Two norms are parallel if an individual holds two roles linked to two independent normative systems, each role belongs to one of these systems and can be satisfied by the same individual physical movement. Note that in this case, there is no behavioral conflict derived from the norms associated to two different roles, as is the case with intrapersonal role conflicts. Now regarding parallel norms it is important to note that even if both norms are satisfied by the same behavior, then each norm corresponds enforcement mechanisms. In other words, the factors that encourage behavior consistent with both norms, the sources of costs, or benefits that produce individual motivations in each case are different.

For example, suppose that according to a constitutional provision in the case of a vacancy in the Supreme Court, the president is required to select the individual who will fill the position and such an individual should hold a law degree and have at least 10 years of experience in the judiciary. Now suppose that the president's best friend satisfies the constitutional requirements, is unemployed, and in great need of work. Now if an interpersonal norm of friendship dictates that one ought to help one's friends if one is in a position to do so, the constitutional provision regulating the selection of Supreme Court Justices and the interpersonal norm in question are parallel norms since they can be satisfied by the same physical movement: the

designation of the president's friend to the Supreme Court vacancy. Now, these parallel norms' enforcement mechanisms are very different: on the one hand, failing to provide help for friends in need would probably inflict costs on the relation, while the cost of failing to satisfy the constitutional requirements for justices' nominations would probably be a failure of confirmation by the Senate. It is noteworthy that a threat of non-compliance with a norm is often enough to activate the enforcement mechanism in an observable way, as was the case in our account of Alemán's failure.

What can we conclude of a case where a constitutional norm has a parallel norm and the only enforcement mechanisms activated by threats of behavior inconsistent with both norms are those of the parallel norm? To follow the previous example, what could we conclude if there are expectations that the president will not nominate his friend but instead someone else who does not satisfy the constitutional requirements to be Justice and Congress signals that would welcome such nomination (while the relation between the president and his friend become distant)? Clearly, if ultimately the president nominates his friend, we cannot claim that the constitutional norm was efficacious even if its requirements were met given that Congress had already signaled that it would not matter if those requirements were not met. If the enforcement mechanisms linked to the constitution did not play any role on the presidential motivation to make such nomination, then we would need to conclude that there was mere text-reality agreement but not constitutional efficacy.

It may be argued that the strategy of focusing on the enforcement mechanisms is not very helpful since it requires a clear threat of constitutional violation or ad hoc amendment. To clarify why this is not necessarily the case take Levinson and Pildes' argument in their article "Separation of Parties not of Powers" (Levinson and Pildes 2006). These authors claim that the United States' system of separation of powers is not efficacious because what motivates members of Congress to limit the executive is fully determined by the dynamics of parties and has little to do with the constitution. To support their claim, they argue that Congress' constitutional mechanisms of enforcement are plagued with collective action problems and, therefore, they are not associated with actual costs for not behaving in agreement with the constitutional norm. According to these authors, party politics are the only source of actual costs for members of Congress. In sum, researchers can design different strategies to study the enforcement mechanisms of constitutional norms and assess through them constitutional efficacy.

The following are cases where enforcement mechanisms are exogenous, but the constitution does play a coordination function by enabling the identification of governmental transgressions. Take Barry Weingast's influential article "The Political Foundations of Democracy and the Rule of Law." Weingast's central question is: "How are democracy's limits enforced?" His aim is to give "a unified approach to the political foundations of limited government, democracy, and the rule of law- phenomena requiring that political officials respect limits on their own behavior" (Weingast 1997, 245). Weingast's approach rests on a game-theoretic model of the stability of limited government that focuses on the relation between a single political official, called the sovereign, and the citizenry. To stay in power, the sovereign requires sufficient support from the citizens, and each individual supports the sovereign as long as he does not transgress what the citizen believes are her rights (Weingast 1997, 246). Different citizens have different "preferences and values" and, therefore, different conceptions of what her rights are (Weingast 1997, 245–6). So accordingly constitutions are devices that *coordinate* the citizens *on* what constitutes a violation of rights so that they can collectively react to transgressions by withdrawing their support from the sovereign. If the constitution functions, that is if citizens are coordinated on its content, the sovereign will avoid any behavior that violates the constitution because by doing so he risks losing power.

Notice that in the model the controls are exogenous to the constitution. Weingast claims that whether or not a constitution coordinates individuals on its content is a function of the social consensus on the rights of citizens and the limits of the state.

> In terms of the model, limits become self-enforcing when citizens hold these limits in high enough esteem that they are willing to defend them by withdrawing support from the sovereign when he attempts to violate these limits. To survive a constitution must have more than philosophical or logical appeal; citizens must be willing to defend it. (Weingast 1997, 251)

> Because citizens have different views about ideal limits, a unique set of ideal limits is unlikely. Coordination requires that citizens compromise their ideal limit...When the difference between each citizen's ideal and the compromise is small relative to the cost of transgression, the compromise makes the citizens better off. (Weingast 1997, 252)

According to this account, whether there is congruence between the constitutional text and the political actor's behavior mainly depends on the presence of a common set of citizen attitudes that are exogenous to the constitution and its incentives. What maintains the equilibrium of text-reality congruence has, therefore, very little to do with the constitution, its roles, and its design. This point is made clear in Weingast's account of why Latin American constitutions "have not worked" while the American has:

> [Latin American constitutions "have not worked" because] Latin American states are not characterized by a common set of citizen attitudes about the appropriate role of government...[While] citizen reaction implies that US constitutional restrictions on officials are self-enforcing ...Latin American states exhibit a complementary set of phenomena: citizens unwilling to defend the constitution, unstable democracy and episodic support for coups. (Weingast 1997, 254)

In sum, in this model the constitution's function is limited to enabling coordination on what constitutes a violation and so that citizens collectively react to the transgression. However, the enforcement mechanisms, the costs that the sovereign knows would suffered if he does not behave in accordance with the constitutional mandates, are exogenous to the constitution, they do not depend on the constitutional roles and powers, and therefore in these cases we cannot claim that the constitution is efficacious.

Finally, we have cases of enforcement mechanisms endogenous to the constitution. In these cases we can affirm the presence of constitutional efficacy strictly speaking. To understand these cases, it helps to distinguish them from the previous ones. As Hardin argues, in claiming that a particular constitution is a device for coordination we could be making two quite different claims. First, we could be claiming that the content of a particular constitution coordinates or coordinated the most important sectors of a society (which are exogenous to the constitution). In other words, that those interests were coordinated *on* the constitution. This understanding of what it means for a constitution to coordinate may be given as an account of a successful constitution-making process, as an explanation of why the content of a particular constitution is such, or as Weingast does, as an account of one of the functions that constitutions have that is serving as focal points (on the functions of constitutions, see Ginsburg and Simpser 2014). For instance, this is the notion that Hardin nicely uses in his account

of the American constitution-making process, which he notes coordinated the most important economic interests and that, we may add following Weingast, also the most important attitudes about the appropriate role of government (i.e. that those interests and attitudes were coordinated *on* the content of the constitution) (Hardin 1998).

Now when we claim that a constitution that is efficacious is a coordination device, we are claiming that actions are successfully coordinated *under* it; that is, that the behavior that is its regulative target is attained, thanks to the incentives the constitution gives to the relevant constitutional role holders. That public actors act according to the constitution as a result of their pursuit of individual benefits under constitutional laws, using their constitutional powers. Paraphrasing Madison's *Federalists 51* an efficacious constitution provides "the personal interests and constitutional means" for its enforcement. In these cases "the interest of the man must be connected with the constitution. . .," (Hamilton et al. 2000) the enforcement mechanisms are therefore endogenous to the constitution, and we can claim that it is efficacious.[10]

Finally, to further clarify how the enforcement mechanisms at play can enable us to assess the efficacy of a constitutional norm, let us return to Alemán's unsuccessful reelection attempt. How can we know whether the 83 Article was efficacious? According to what I have argued, we need to analyze the enforcement mechanisms and their relation to the constitution. In particular, we need to assess whether the enforcement mechanisms were parallel, exogenous, or endogenous to the 1917 Constitution. I have already identified the two enforcement mechanisms that were at play in Alemán's failure: those linked to the intra-party opposition to Alemán's reelection led by President Cárdenas, President Ávila Camacho, and General Taboada and those associated with the opposition formed by former members of the PRI who had decided to split from it as a response to Alemán's attempt.

We know that in Alemán's succession there was text-reality agreement since he was not reelected. Now, I believe the account I have provided of the case shows that the enforcement mechanisms that enabled such agreement were not endogenous: the crucial enforcers (Cárdenas, Ávila Camacho and Taboada) did not hold at the time any constitutional role[11] and the costs they could infringe over Alemán were independent of the constitutional functions or powers. Therefore, we can conclude that according to my account, this is not a case of strict constitutional efficacy.

However, as is evident also in the account, the 1917 Constitution played an important function in the enforcement of Article 83. All enforcers coordinated *on* its content: "The citizen who had performed as President

of the Republic... *in no case and under no circumstances* may perform again this position" (Art. 83 1917 Constitution). The constitution was an ever-present reference, and as Cosío Villegas nicely puts it, there was a "serious danger that Cárdenas and other great personalities would decide to move the opposition to the terrain of arms, and they would have an excellent flag to make a military movement succeed" (Cosío Villegas 1975, 120): the violation of the 1917 Constitution.

In sum, an analysis of the enforcement mechanisms at play in Alemán's failure enables us to conclude that while in this case we cannot assert constitutional efficacy, we can say that the constitution function as a device *on* which enforcers were coordinated. Hence, *pace* Weingast, this Latin American constitution "functioned" according to his model but under a nondemocratic regime.

Conclusion

I analyzed a case of constitutional enforcement in autocracies. I presented an account of why President Alemán failed to violate or amend Article 83 of the 1917 Mexican Constitution to enable his reelection, even if he was a president with extraordinary power in a nondemocratic regime. I discussed whether and to what extent was Article 83 efficacious in this case. Furthermore, I argued that this account illuminates how to assess constitutional efficacy more generally, and hence how can we respond to the challenge posed by observational equivalence of different types of motivations to behave in accordance with the constitution. I claimed that to determine the degree of constitutional efficacy of a constitutional norm it is necessary to identify the mechanisms of constitutional enforcement and their relation to the constitution. It is noteworthy that if my account is correct, some of the functions that have usually been ascribed to constitutions in democratic contexts, such as being a coordination device on which enforcers coordinate, are common to constitutions in certain authoritarian regimes (on this point see Ginsburg y Simpser 2009).

An important question that naturally derives from the account presented and that is not answered here is: Can strict constitutional efficacy be attained in nondemocratic regimes? I do not answer this question here because (similarly to the argument regarding the claim that autocrats can always make ad hoc amendments) to understand the roles of constitutions in autocracies, and its differences from those in democracies, it is important to proceed from empirical studies, and not from *a priori* preconceptions of how "all" autocratic regimes work.

Notes

1. For instance, the constitution mandated life tenure for Supreme Court Justices. However, every 6 years the incoming president used to appoint as much as 72% of the Court (Ruiz Cortines, 1952–1958) and no less than 36% (López Mateos, 1958–1964). "The president could thus somehow create vacancies to be filled by justices he appointed or, in other words, he could either dismiss justices or induce early retirements" (Magaloni 2003, 228–289). See also: Valdés Ugalde (2010).
2. Every incoming president amended the constitution to make it fit his political agenda: as much as 66 constitutional provisions were altered in the presidential term of Miguel de la Madrid Hurtado (1982–1988).
3. An example of these changes was the transformation of the selection of candidates from primary elections to local party assemblies that enabled more control of the party leaders over the governors, senators, and deputies candidacies (Servín 2001, 129).
4. The PRI had a formal president but as stated earlier the President of the Republic was the political leader of the party.
5. As I have said Lombardo Toledano was an important union leader who strongly opposed Alemán's reelection. He was close both to President Cárdenas and to President Ávila Camacho.
6. For instance, many messages were expressed in negative form. As I stated before, the Generals' denial of the "rumors" of a possible split from the PRI left wing actually brought that possibility to the table, and this was the way the message was understood by the politicians of the time and by the foreign analysts. In the same connection, stating that the president *would* never promote his reelection actually meant that he *shouldn't*.
7. General Sánchez Taboada (1895–1955) was a hero of the Constitutionalist Army. He executed the death sentence of Emiliano Zapata. He was Governor of Baja California, Secretary of Marine, and president of the PRI both in Mexico City and at national level. It was known that General Taboada supported the presidential candidacy of Fernando Casas and strongly opposed any attempt to amend the constitution.
8. He eventually deposed his candidacy in favor of General Henríquez to more effectively "defend the principles of the revolution."
9. Note this conceptualization of constitutional efficacy refers only to the organic sections of constitutions (i.e. to articles that establish the functions and powers of constituted organs).
10. Note that the need of separating these two senses in which a constitution is a coordination device follows from the recognition that an account of modern constitutional government requires a two-stage theory (see Hardin 1998, 83).
11. Neither ex-president nor PRI president has constitutional status (i.e. they are not part of the constitution).

References

Alemán Valdes, Miguel. 1987. *Remembranzas y testimonios*. Cd.Mx.: Grijalbo.

Angulo et al. 1950. Desplegado, *El Universal*, (Mexico City), April 15.

Barquín Alvarez, Manuel. 1990. Artículo 83. In *Constitución política de los Estados Unidos Mexicanos: comentada*. Cd.Mx.: IIJ-UNAM.

Barros, Robert. 2002. *Constitutionalism and Dictatorship*. Cambridge: Cambridge University Press.

Cárdenas, Lázaro. 1973. *Obras I. Apuntes, 1941–1956*. México City: Universidad Nacional Autónoma de México, v. 2.

Chanes Nieto, José. 1993. *La designación del presidente de la república*. México: Plaza y Valdés Editores.

Cosío Villegas, Daniel. 1975. *La Sucesión Presidencial*. Cd.Mx.: Cuadernos de Joaquín Mortiz.

Fix-Fierro, Héctor, and Diego Valadés. 2016. *La Constitución Reordenada y Consolidada*. Cd.Mx.: Instituto de Investigaciones Jurídicas UNAM. Available online in http://www2.juridicas.unam.mx/constitucion-reordenadaconsolidada/

Ginsburg, Tom, and Alberto Simpser, eds. 2009. *Constitutions in Authoritarian Regimes*. New York: Cambridge University Press.

Ginsburg, Tom, and Alberto Simpser, eds. 2014. *Constitutions in Authoritarian Regimes*. New York: Cambridge University Press.

González Casanova, Pablo (coord.). 1985. *Las elecciones en México: Evolución y Perspectivas*. Cd.Mx.: Siglo XXI.

Hamilton, Alexander, John Jay, and James Madison. 2000. *The Federalist*. New York: The Modern Library.

Hardin, Russell. 1998. *Liberalism, Constitutionalism, and Democracy*. New York: Oxford University Press.

Levinson, Daryl J., and Richard H. Pildes. 2006. Separation of Parties, Not Powers. *Harvard Law Review* 119 (8): 2312–2385.

Magaloni, Beatriz. 2003. Authoritarianism, Democracy and the Supreme Court: Horizontal Exchange and the Rule of Law in Mexico. In *Democratic Accountability in Latin America*, ed. Scott Mainwaring and Christopher Welna. New York: Oxford University Press.

Marván Laborde, Ignacio. 2017. *Cómo hicieron la Constitución de 1917*. Biblioteca Mexicana, Cd.Mx.: Secretaría de Cultura-Fondo de Cultura Económica-CIDE.

Medin, Tzvi. 1990. *El sexenio alemanista*. Cd.Mx.: Era.

Medina, Luis. 1982. Civilismo y modernización del autoritarismo. In *Historia de la Revolución Mexicana, 1940–1952*, vol. t.20. Cd.Mx.: El Colegio de México.

Mendieta y Núñez, Lucio. 1949. Irresponsable Goverments. El Universal November 26, 1949.

Merton, K. Robert. 1949. *Social Theory and Social Structure*. New York: The Free Press.

Pozas-Horcasitas, Ricardo. 2009. Elección presidencial y reproducción del régimen político en 1964. In *Secuencia*, Núm. 74, Mayo-Ago.

Pozas-Loyo, Andrea. 2012. Constitutional Efficacy. PhD Dissertation, New York University.

Pozas-Loyo, Andrea and Julio Ríos-Figueroa. 2017 (forthcoming). Authoritarian Constitutionalism. In *Oxford Handbook of Constitutional Law in Latin America*, ed. Roberto Gargarella and Conrado Hübner Mendes. New York: OUP.

Searle, John R. 2010. *Making the Social World*. New York: Oxford University Press.

Servín, Elisa. 2001. *Ruptura y oposición: el movimiento henriquista, 1945–1954*. Cd.Mx.: Cal y Arena.

Servín, Elisa. 2002. Las elecciones presidenciales de 1952: Un intento de cambio democrático. *Estudios de Historia Moderna y Contemporánea de México* 23 (23): 179–205.

Torres, Blanca. 1984. Hacia la utopía industrial. In *Historia de la Revolución Mexicana, 1940–1952*, vol. t.21. Cd.Mx.: El Colegio de México.

Tushnet, Mark. 2015. Authoritarian Constitutionalism. *Cornell Law Review* 100: 391–460.

Valdés Ugalde, Francisco. 2010. *La regla ausente. Democracia y conflicto constitucional en México*. Barcelona: Editorial Gedisa.

Weingast, B.R. 1997. The Political Foundations of Democracy and the Rule of the Law. *American Political Science Review* 91 (2): 245–263.

Wilkie W. James, and Edna Monzón. 1969. *México Visto en el Siglo XX*. Cd.Mx.: Instituto de Investigaciones Económicas UNAM.

"Führer befiehl, wir folgen dir!" Charismatic Leaders in Extremist Groups

Michael Baurmann, Gregor Betz, and Rainer Cramm

Russell Hardin's Economic Theory of Knowledge

The economic approach to explaining individual behavior has undergone significant changes and enhancements in the last decades. Traditional rational choice is based on the presupposition of given preferences which, in the face of external restrictions and on the basis of subjective beliefs, are translated into action by rational decisions. The assumption that we can explain the behavior of people in general as a result of optimizing rational choices was contested already quite early by the theory of bounded rationality. Since then the overwhelming empirical findings of countless experimental and field studies have proved conclusively that people in their actual behavior practically never meet the rigorous requirements of standard rational choice theory.

The questioning of the presupposition of homogeneous and stable preferences does not go back so far as the attack on the assumption of rationality. But in the meanwhile, it is also part of a more or less mainstream

M. Baurmann (✉) • R. Cramm
Institute of Social Sciences, Heinrich-Heine-University Dusseldorf, Düsseldorf, Germany

G. Betz
Institute of Philosophy, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_10

criticism on rational choice theory to stress the empirical evidence we have for the heterogeneity of preferences, for example, in regard to altruistic and retributive preferences, the adaptation of aspiration levels to feasible opportunities, or the phenomenon that intrinsic motivation can be crowded out or reinforced due to contextual factors.

But one cornerstone of traditional rational choice had received amazingly little critical attention until Russell Hardin published his "How Do You Know" in 2009. The question how we can integrate an empirically convincing explanation of belief formation into a rational actor theory and sort out the role different kinds of beliefs play as motivating factors for human actions was not on the agenda of important research desiderata. This is somewhat astonishing as our beliefs about the facts in the world or the importance of certain values and norms are obviously decisive for our way of acting. Therefore the empirical processes by which we acquire these beliefs should have been of utmost interest for every theory of action.

In the case of normative beliefs, the neglect is maybe an even more serious omission, induced by the erroneous assumption that normative beliefs are just "cheap ideas" that have no real influence on human behavior. But, as Russell Hardin stresses straight at the outset of his book, we have to acknowledge that moral or religious principles come to many people as facts "no different in kind from other facts, such as the moon goes through its various phases" (Hardin 2009, 18). This kind of everyday objectivism does not only open up the possibility that people act according to their moral or religious beliefs just as regards their descriptive beliefs, but that they may adopt moral or religious beliefs that lead them to act in ways that are against their genuine interests (cf. 17)—a possibility that must be as irritating as it is fascinating for a rational actor theory.

Russell Hardin proposes an economic theory of knowledge as an approach to closing this gap in rational actor theory. The theory is economic in the sense that it strives to explain the knowledge base of average persons as being the result of choices in which people weigh up the costs and benefits of gaining certain pieces of knowledge (cf. 2ff.). Such a theory understands the acquisition of knowledge as an essentially rational process of considering the trade-offs between the value of any kind of knowledge and the value of other things which compete with the investment in knowledge acquisition: "The theory would not be about what the philosophical epistemologist's criteria for truth claims should be, but rather why we come to know what we know or believe" (xi). As the criteria ordinary persons apply to judging their beliefs "are not necessarily criteria for truth, but

merely and genuinely criteria of usefulness" much of the knowledge people accept and act on will be "merely satisficing knowledge, that is, good enough" (24f.). If we want to understand human behavior in this area what is required, therefore, is not a philosophically general theory of knowledge but a "street-level account" (Hardin 1992), a pragmatic theory that focuses on the actual ways people come to hold their beliefs but that bears little resemblance to the "theories of knowledge of those in ivory towers" (Hardin 2009, 19).

Implicated in this approach is a further deviation from philosophical epistemology by using a very broad concept of knowledge which follows the everyday use of this term, and makes no general distinction between beliefs and knowledge or between moral and factual knowledge. An economic theory of knowledge aims at including a vast area of various kinds of belief and behavior, "such as ordinary moral choice, religious belief and practice, political participation, liberalism, extremism, popular understandings of science, and cultural commitments" (3).

Russell Hardin's approach exhibits a family resemblance with social epistemology as it starts from the same basic and almost trivial fact that nearly all of our information and knowledge is not gained by our own experience, investigation, and deliberation but via testimony. Most of an individual's knowledge is socially generated and a result of a division of labor in the production of knowledge (cf. 5). We have no other option than to rely on others if we want to participate in the collective knowledge of our world. Both theories emphasize in this context the important role of epistemic authorities and experts. Contrary to scientific knowledge, ordinary knowledge "is almost entirely grounded in hearsay from a supposedly credible or even authoritative source" (1). So "we first have to judge a particular authority, and then we infer the truth of the authority's claim" (11). Hardin suggests that this deference to authority may be also essential in moral judgments as it "is only an extension of normal reasoning to let specialists assess religious matters and moral matters of right and wrong" (15).

But, in contrast to social epistemology, Russell Hardin is not interested in the question whether and under what conditions information via testimony could create "justified true beliefs". He is interested in the question how people in fact gain information and knowledge. An economic theory of knowledge is an empirical theory of epistemic processes, not a normative theory. However, as an economic theory of knowledge is an offspring of rational choice theory, it could at least be judged as weakly normative in that

it looks for a rational reconstruction of the factual processes of belief formation.

The "Crippled Epistemology" of Extremism

One field to which Russell Hardin applies an economic theory of knowledge is the phenomenon of extremist beliefs (cf. 185ff.). At first sight this may appear as a quite unusual subject for a theory of knowledge. In our paper we want to demonstrate the fruitfulness of this approach and—inspired by Russell Hardin's pioneer work—to describe and analyze a social-epistemic mechanism that can help to explain the emergence, stability, and erosion of extremist opinions in a group.

We thereby share two basic assumptions with Russell Hardin. First, that the acquisition of extremist beliefs follows the same patterns and processes as the acquisition of beliefs about the facts in the natural or social world. People come to believe the truth of extremist world views in the same way as they come to believe the truth of physics or the weather forecast. And we also agree that it is a "crucial move" for an explanation of extremist thinking when we recognize that people learn extremist ideas the same way they learn other things (cf. 159). Second, as the acquisition of most of our beliefs is to be explained as social and not as individual processes, this also applies to extremist beliefs. Hardin's general claim, already noted above, that in the course of these social processes people may adopt moral or religious beliefs that lead them to act in ways that are not in their interest is of special relevance when dealing with extremist or other deviant convictions.

Russell Hardin presents his approach as a serious alternative to psychological or traditional sociological explanations. He argues that we should analyze the dynamics of extremist thinking in groups as a social-epistemic process on the collective level and not as a process that can be attributed primarily to individuals and their idiosyncrasies: "It is generally the group that produces and sustains fanaticism" (185). That does not mean that Hardin abandons an individualistic methodology, but rather that we should understand the formation of individual convictions and opinions as a complex result of multifaceted interactions of people in their social networks and relations. Of course, how much variance such an epistemic approach actually could explain in this difficult and heterogeneous field is ultimately an empirical question.

If a social-epistemic process on the group level is crucial for the emergence and consolidation of extremist beliefs, it is essential to know the

special characteristics of groups in which extremist thinking can flourish. Hardin focuses on three factors which he summarizes as "the crippled epistemology of extremism" (Hardin 2002). The first factor is the inflicted or self-chosen isolation of a group of like-minded people by which the beliefs of its members are constantly reaffirmed and may become more and more polarized. This can work even though for the overwhelming majority of other people outside the group these beliefs sound bizarre and absurd. The second factor is an effective norm of exclusion by which the less intensely committed members of a group and the moderates exit while the most dedicated and extremist remain. The third factor is the crucial role of epistemic authorities in propagating and transmitting extremist views in a group and the unconditional devotion of the group members to their ideological and political leaders.

Hardin summarizes the conditions for a crippled epistemology of groups: "If I am in a small community holding beliefs that others outside that community would think very odd, I may find those beliefs not at all odd because, after all, they are held by everyone I know. They may be merely part of the vast catalog of beliefs that I hold from dependence on authority" (Hardin 2009, 187).

Charismatic Leaders in Extremist Groups

The empirical evidence supports Hardin's analysis. Especially the impressive studies on religious fundamentalism of the "The Fundamentalism Project" (Chicago 1987–1995) which was directed by Martin E. Marty and R. Scott Appleby shows convincingly that groups can develop an idiosyncratic "enclave culture" which is successfully isolated from external influences and that the impact of "charismatic" leaders as ideological authorities is decisive in practically all groups for the inculcation and maintenance of fundamentalist world views.

The crucial role of charismatic leadership is especially salient for religious fundamentalism because the "holy texts" such as the Bible, the Thora, or the Koran reveal their alleged fundamentalist messages not without a heavily biased and selective interpretation. And in most religious traditions the interpretation of holy texts is the exclusive task of religious authorities who make the mission of these texts comprehensible for the ordinary believer and religious layperson. But as heretic religious groups do not recognize the official authorities of their institutionalized and "secularized" denominations, religious authority and leadership in these groups come into

being through an attribution of charismatic qualities to certain persons by the members of the group themselves. The ascription of extraordinary abilities, religious virtuosity, exceptional leadership and moral virtues is the basis for the enthronement of omnipotent religious and political authorities who are in a position to induce extremist and fundamentalist convictions among their followers (cf. Baurmann 2007, 2010a).

But we cannot be content with just stating the fact that the formation of certain variants of extremist groups is regularly dependent on the existence of charismatic leaders. The existence of a superior authority in a group is one possible explanatory factor; however, the emergence of such an authority *is in need of explanation itself*. Leadership does not operate in a vacuum but must be based on a group of potential followers who can be convinced and mobilized. The "charisma" of persons is therefore not a self-evident cause of their exceptional authority. It has to be clarified instead which social conditions and processes in a group lead to the attribution of a special "charisma" to certain persons so that they are established as supreme ideological leaders whose epistemic authority is so potent that they are able to generate devoted followers and convert them to radical believers that are normally rejected by the large majority of the surrounding society.

We can characterize this sovereign position of power as a position in which a person enjoys *exclusive epistemic trust* of the group members. This trust must be accompanied by a corresponding strict mistrust toward all people outside the group and toward competing epistemic authorities who on no account are to be accepted as alternative sources of information and knowledge. The emergence and consolidation of charismatic leadership in a group is necessarily combined with the formation of a group-specific *particularistic trust*—in the social as well as in the epistemic dimension.[1]

Epistemic trust includes social trust in the personal integrity and benevolence of persons and, in addition, confidence in their special competence and cognitive faculties which together can motivate others to accept and adopt their opinions and views. In the case of "charismatic" authorities, this can imply indoctrinating their followers with ideologies and convictions that differ significantly from their initial belief systems and world views: "it is written, but I tell you!" But even charismatic leaders cannot develop their messages in an empty space. They must connect with what is—already—*written* and present in the life world of their addressees. The more they manage to do this, the more plausible their message will appear and the less they have to utilize their "capital" of charisma to convince their followers. Therefore we have to take into account that the evolvement of radical and

extremist ideologies in a group will often be an incremental process in which the faith in certain leaders and the adoption of their views will develop mutually and gradually in a self-reinforcing dynamic.

Our central explanandum then is: *How can exclusive epistemic trust in a certain person evolve and stabilize in a group so that this person is able to implant and disseminate extremist and deviant views among the group members?*

A Social Mechanism of Opinion Dynamics

This process can be explained if we understand the underlying social mechanism. We assume that this mechanism is a special case of a social–doxactic mechanism which determines opinion dynamics in social groups in general (cf. Baurmann et al. 2014). The core of this mechanism is constituted by a process of mutual influence and adaptation in which individual experiences and deliberations are continuously compared and adjusted in accordance with the experiences and deliberations of other persons who are considered relevant and reliable. In detail we make the following assumptions:

1. Persons influence each other mutually in their opinions on the basis of *epistemic trust*. The greater the epistemic trust in a person, the more other people will orient themselves according to the opinions of this person.
2. Epistemic trustworthiness is based on *coherence*, *competence*, and *veracity*. *Coherence* means that the opinions of another person must appear plausible to be taken seriously, they should not diverge too much from one's own already established opinions but have to stay within a certain confidence interval or opinion space. *Competence* refers to the ability of a person to acquire reliable knowledge and sound insights in a certain area. *Veracity* is attributed if it is assumed that the incentives of the social context and the motivational dispositions of persons will lead them to transmit their knowledge and insights truthfully to their recipients.
3. *Epistemic self-confidence* is based on the competence persons ascribe to themselves. The lower the epistemic self-confidence of persons, the more they will be inclined to adapt to the opinions of other people who they judge to be epistemically trustworthy.
4. Opinion formation involves *first-order opinions* about the issues that are relevant in a certain field and *second-order opinions* about the

epistemic trustworthiness of persons who express their opinions about these issues. First-order opinions can include descriptive as well as normative opinions. Second-order opinions refer to characteristics of persons that are relevant for their quality as epistemic sources.

5. Persons *influence each other mutually* both in the formation of their first-order opinions and their second-order opinions. They consider the opinions of other trustworthy persons with regard to descriptive and normative issues as well as with regard to their estimation who is competent and reliable to pass considered judgments over these issues.

As noted above, these factors constitute a general socio-doxastic mechanism and as such do not signal any "abnormalities". Our central research hypothesis suggests that the emergence of extremist opinions in certain groups is the result of the nuts and bolts of this general mechanism and of the predominance of external conditions that constitute a deficient epistemic environment, much in the sense of Russell Hardin's crippled epistemology—meaning not as a result of psychology, irrationality, or individual deviance. To put it pointedly, one can become an extremist because one lives in a pathological epistemic environment and not because of a pathological personality (cf. Baurmann 2007).

It is an important feature of the described mechanism that it not only explains the group-induced development of first- and second-order opinions but that it also depicts the dynamic relationships between these different layers of opinion formation. On account of this structure, persons will be influenced by other persons not only in regard to their opinions about political options, societal connections, or ideological world views. This adaptation process itself will in turn be intertwined with the mutual adaptation of the second-order opinions about who has sufficient or special competence to understand and judge such options, connections, or world views. These two-layer dynamics could result in far-reaching transitions of the initial convictions of persons so that they ultimately may adopt extremist and radical opinions which were originally not within their opinion space and may well have appeared absurd to them.

We think that precisely in the interrelations between opinions of the first and second order lies the key to an explanation of how it can come about that even in a group in which initially neither an outstanding leader was generally accepted nor extremist views were held by the majority, a development can take place that finally leads to the establishment of an

uncontested ideological leader under whose influence all other group members adopt convictions which differ drastically from their original world views.

But how such a mechanism works exactly and how the different factors affect its mode of action in detail are open questions. They are not to be answered easily, not least because the postulated mechanism exhibits a considerable internal complexity due to its multi-level structure. It is not possible to analytically determine the results of opinion dynamics in a group with many members after prolonged sequences of mutual influence on different levels or the impact smaller or larger changes of individual parameters or external conditions will produce. On the other hand, the basic elements of the supposed mechanism and their fundamental interrelations are quite simple. The challenges for analyses only begin when we have to deal with interrelations involving large numbers of actors over long periods.

Mechanisms of this kind, therefore, are predestinated for experimental simulations. In the following we want to show how on the basis of an idealized mathematical model some of the fundamental aspects of the relevant dynamics could be explored with such simulations. These models and simulations could not themselves deliver explanations and they cannot substitute an empirical examination of theories. But they are potentially powerful instruments to develop new and fruitful hypotheses in a systematic and transparent way. They could help to illuminate the complexity of social dynamics and to detect concealed and analytically incomprehensible consequences of theoretical assumptions (cf. Hegselmann and Flache 1998).[2]

Simulation of Opinion Dynamics in Extremist Groups

Structure of the Simulation Model

We have developed a simple prototype of a simulation model for opinion dynamics which provides promising first results (cf. Baurmann et al. 2014). The basic factors and relations which, according to our assumptions, are constitutive for the general social mechanism of opinion dynamics are operationalized in the model as follows[3]:

1. The model describes how the opinions of n agents change in the course of time (discrete time steps, $t = 0,1,2,3, \ldots$).
2. Each agent possesses a first-order opinion which is represented by a real number between 0 and 1.

3. Each agent assigns himself and the other agents degrees of epistemic competence on a scale between 0 and 1. Accordingly each agent possesses n second-order opinions.
4. An agent A trusts another agent B iff (i) B's first-order opinion are inside the confidence interval of A and (ii) A assigns according to his second-order opinions to B at least the same level of competence as to himself.[4]
5. *First dynamic principle*: the first-order opinions of an agent A at time step $t+1$ equals the average of the first-order opinions of all agents at time step t whom A trusts at t.
6. *Second dynamic principle*: the second-order opinion of agent A about the degree of competence of B at time step $t+1$ equals the average of the corresponding opinions of all agents at time step t whom A trusts at t.

Because the model abstracts from all other factors which influence our opinion formation as well (sympathy, argumentation, complexity, interests, emotions, etc.), it is a strongly simplified reconstruction of an in fact highly complex process. The model, therefore, is neither suitable for complete explanations nor prognostic aims (cf. Betz 2006, 2010). But, on the other hand, exactly because of its idealizations the model facilitates examination of the special aspects which are under consideration here with high precision and particularly rigorously. This will contribute to the heuristic value and explanatory potential of the hypotheses which are deducible from the model.

The outcomes will relate especially to the intertwined dynamics of first- and second-order opinions. With their help we can generate hypotheses about how it is possible that persons with extremist opinions can accumulate the necessary exclusive epistemic trust in a group to become a "charismatic" leader and in this way successfully disseminate extremist opinions that were initially outside the horizon of the other group members. Of course, whether such a process of mutual adaptation of first- and second-order beliefs in fact plays an important or maybe even decisive role in the emergence and dissemination of extremist world views can only be clarified by empirical studies.

Our model combines and extends the Lehrer-Wagner model (Lehrer and Wagner 1981) on the one hand and the Hegselmann-Krause-model (Hegselmann and Krause 2002, 2006; Hegselmann 2004) on the other hand. In both models beliefs are represented by real numbers in the unit

interval. With the Lehrer-Wagner model, we share the idea that the involved persons mutually ascribe to each other different degrees of competence (second-order opinions).[5] But, as in the Hegselmann-Krause-model, the new beliefs of a person are not a result of just a weighted average but are subject to the bounded-confidence mechanism, respectively the coherence restriction. Particularly the inclusion of variable second-order opinions differentiates our approach from previous models.[6] This innovative element allows the reproduction and simulation of much more complex opinion dynamics than the alternative models.

We also think that interrelations between first- and second-order opinions are in fact an essential part of the empirically observable opinion formation processes. If this is the case, then simulation models should include this structure because these models should not only reproduce end states that are compatible with empirical facts but should also aim at reconstructing the causal mechanisms as adequately as possible (cf. Hedström and Swedberg 1998; Hedström and Ylikoski 2010).

A precise formal description of our model can be found in Appendix 1.

First Experiment: Emergence of Extremist Groups

As already stated, the ideological power of charismatic leaders is based on the exclusive epistemic trust of their followers which corresponds to a correlative mistrust toward all other epistemic sources and authorities. The findings of the "Fundamentalism Project" prove that all studied groups indeed make great efforts to secure particularistic in-group trust and social isolation and immunize their ideology against alternative world views and divergent experiences and influences. These strategies aim at ensuring that the group members will not develop any reliance on persons who do not belong to their own group.

Two simulation experiments with our model support the assumption that the absence or rather the undermining of external trust relations is just as crucial for the formation as for the stabilization of extremist groups and their internal hierarchical structure with a "charismatic" leader.

We analyze a group with 10 persons as members. Persons P2–P10 have moderate first-order opinions (0.5, 0.55, and 0.6); only person P1 takes an extreme position with a first-order opinion 0.9. The confidence interval of all persons is 0.33. The extreme position of P1 is compatible only with the confidence interval of P2 who holds the first-order opinion 0.6. The initial trust relations are depicted in Fig. 1a. Persons P5–P10 trust each other

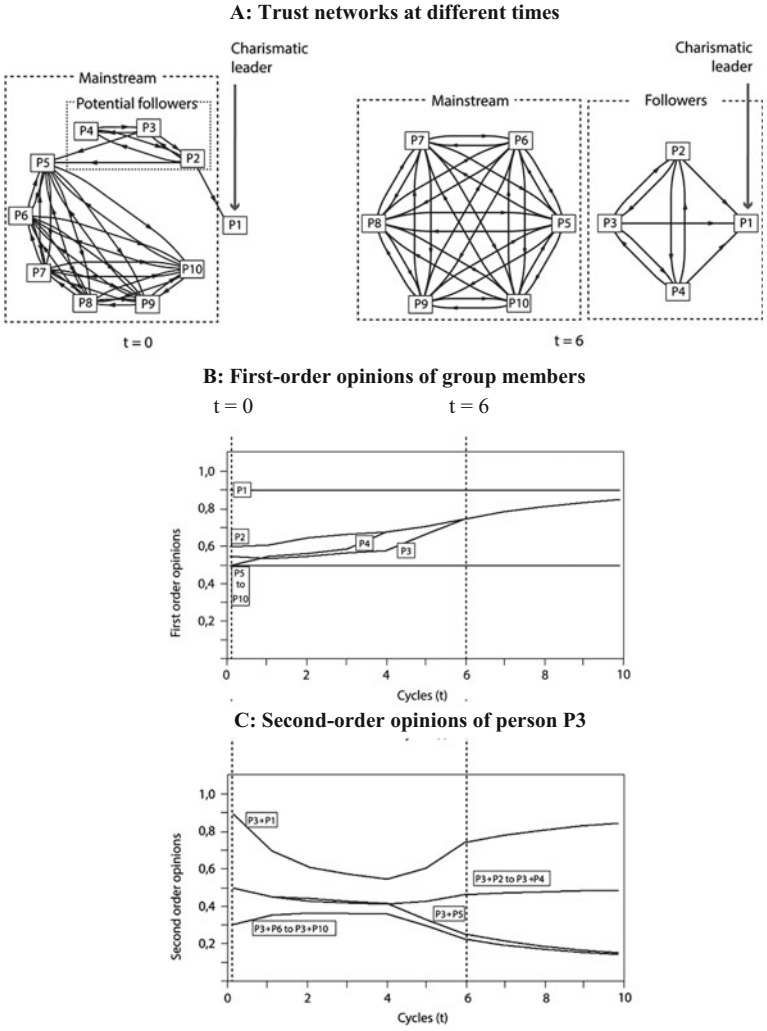
mutually. The same applies to persons P2-P4. P2 and P3 in addition trust P5, and P2 also trusts P1. But P1 trusts nobody except himself. This maximal level of epistemic self-confidence is an important precondition for becoming a group-leader who is able to impose his personal opinions on the group. Because P1 does not concede to any other person the same degree of competence as himself, his self-confidence can neither be shaken by divergent second-order opinions of other persons nor will his extreme first-order opinions be challenged by more moderate views in his environment.

These assumptions model a situation in which a group (P1–P4) already experiences a significant degree of social isolation. Due to their thin epistemic trust relations to persons outside their group their opinion formation is largely shielded against influences from outside. Therefore important preconditions for a "crippled epistemology" are fulfilled.

If we run a simulation of the opinion-formation process starting from this situation, already after a few steps a group evolves which is characterized by extreme opinions and an exclusive epistemic trust toward a charismatic leader who is the source of the progressive dissemination of these opinions in the group.

The first-order opinions of P2-P4 continuously adjust to the extreme position of P1 (cf. Fig. 1B). The initial trust relations of P2 and P3 with P5 (Fig. 1A, $t = 0$) are broken off step by step, and whereas P2 has trusted extremist P1 from the beginning, P3 and P4 follow him in steps 4 and 5 and also develop trust toward P1. P1 consequently becomes an uncontested authority (Fig. 1A, $t = 6$) who can impose his own extremist views without compromise on his new followers. It is noteworthy that by this process P3 and P4 accept an extremist position in the end, even though this position was outside their confidence interval at the beginning and must have appeared distinctly "implausible" to them because of the incompatibility with their already established beliefs. Decisive for the development of the extremist group is therefore the "intermediary" P2 who radicalizes the opinions of P3 and P4 at first only moderately until they finally enter the sphere of influence of P1.

Figure 1C demonstrates how second-order opinions play an essential role in these dynamics. It shows the development of the second-order opinions of P3. Up to the fifth step P3 judges P5, a member of the mainstream group, as at least as competent as herself. Consequently the trust relation to P5 stays intact and the subgroup of P3 is not yet completely isolated. But in the fifth step, P3 for the first time develops trust in the



Numeric specification in Appendix 2.

Fig. 1 Simulation of the emergence of extremist groups, sufficient exclusivity of trust relations

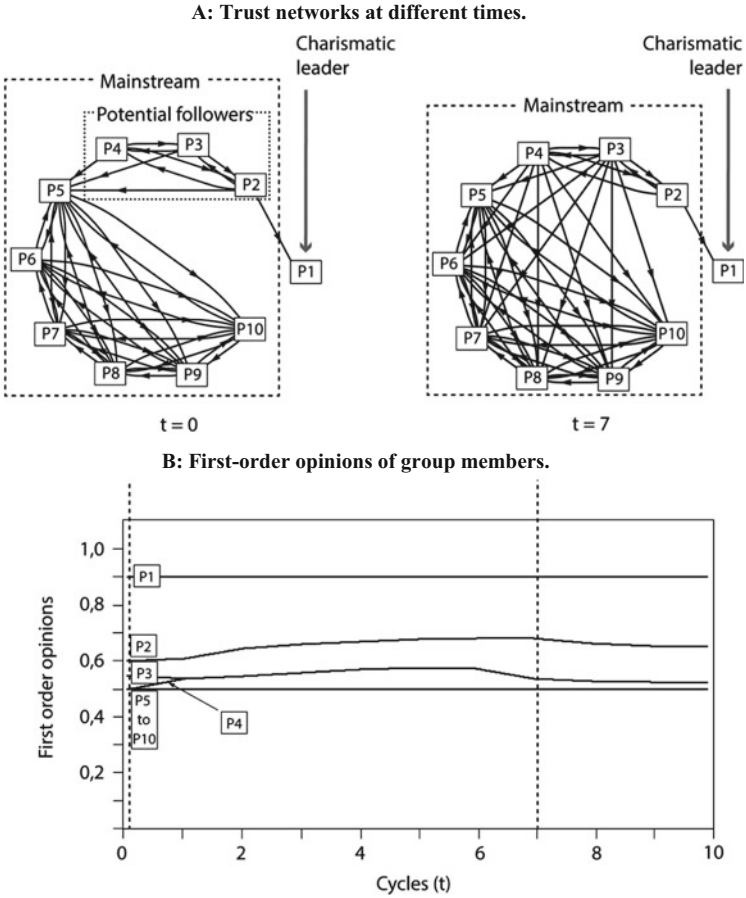
potential leader P1 whose second-order opinions about P5 (0.1) therefore become relevant to P3. As a result P3 attributes lower competence values to P5. And already in the sixth step, P3 does not trust P5 any longer. The group is completely isolated (Fig. 1A, $t = 6$).

To study the relevance of exclusive trust further, we minimally vary the virtual experimental design. We assume that not only P2 and P3 initially trust P5, but that P5 is also trusted by P4. In contrast to the former initial conditions, P4 now judges P5 as slightly more competent (higher second-order opinion). Apart from that all other conditions remain identical. The resulting trust relations at $t = 0$ are represented in Fig. 2a.

The simulation of the opinion dynamics in this case results in a completely different picture although the starting conditions appear quite similar: no extremist group evolves. The additional trust relationship between the potential followers of P1 and the mainstream prevents the recognition of P1 as a charismatic leader (Fig. 2A, $t = 7$). Instead of breaking off their relations to the mainstream, P2–P4 extend them in fact. Moreover, P3 and P4 only temporarily develop trust in P1 (step 6). But as they deepen their trust relations to the mainstream, at the same time, the extremist P1 is already in step 7 no longer within the limits of the confidence interval of P3 and P4. Only P2 continues to trust P1 and positions herself eventually between the poles of the extremist P1 on the one hand and the mainstream on the other hand—with a bias toward the mainstream because P2 trusts more than one person there.

The first experiment corroborates the theoretical and empirical conjecture that exclusive epistemic trust in an opinion leader could be a crucial explanatory factor for the emergence and dissemination of extremist convictions in a group. The correspondence between the results of the experiment and the facts that are known about extremist groups could be deemed as an indicator of the adequacy and heuristic potential of the simulation model.

But the simulation does not only elucidate how the influence of charismatic leaders could determine the convictions of all other members of their groups. It also emulates the opinion dynamics by which in a stepwise transition of the trust relations in a group such a leading figure is established in the first place. This was the explanatory task we postulated: *By what social mechanism can exclusive epistemic trust in a certain person evolve in a group and establish that person as an uncontested epistemic authority?* In the simulation model such a mechanism is driven by the intricate interrelations between opinions of first and second order. The establishment of a



Numeric specification in Appendix 2.

Fig. 2 Simulation of the emergence of extremist groups, insufficient exclusivity of trust relations

charismatic leader is the result of a mutual adaptation of the judgments of group members as to which persons are epistemically and socially trustworthy and which persons have to be regarded with suspicion. The results of the simulations, therefore, support the hypothesis that an explanatory approach that is based on the relationships between first- and second-order opinions

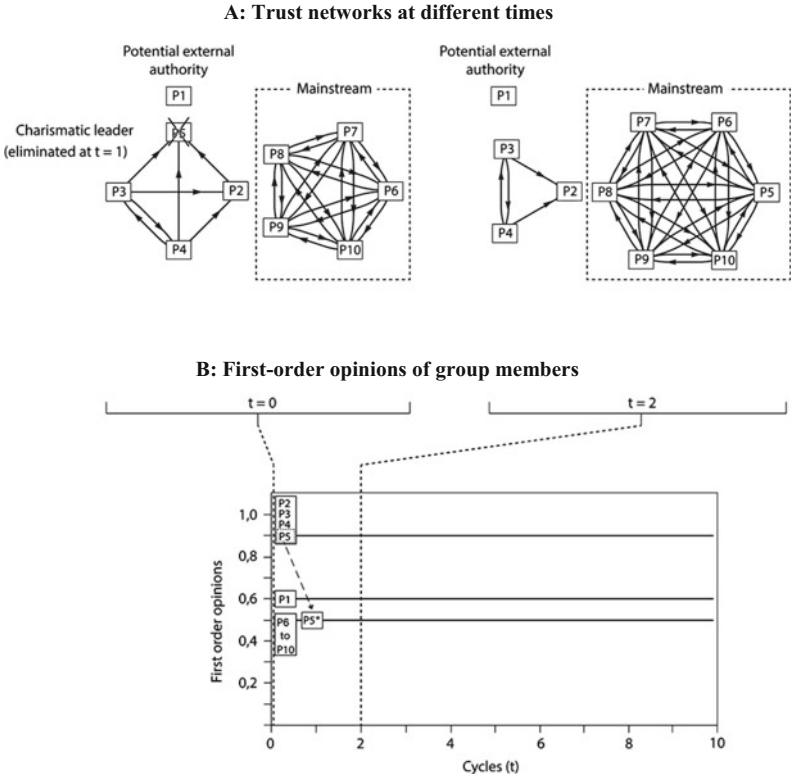
could be particularly promising in explaining the emergence of extremist groups and the enthronement of their ideological leaders.

Second Experiment: Stability of Extremist Groups

In contrast to the previous case, in our second simulation experiment we are not studying the emergence but the stability of extremist groups. We start with a situation in which an extremist group already exists. It is a situation in equilibrium which means that without external influence there would be no change in the opinion structure and the group would remain stable. In this initial situation, besides the extremist group (P2–P5), there is a mainstream group (P6–10) and a loner (P1) whose opinions are less radical than the opinions of the extremists. The confidence interval of all persons is 0.25. Between the three factions no trust whatsoever prevails, as the extremist group has successfully cut off all external trust relations. P5 is the charismatic leader of the extremists. He only trusts himself, whereas he is trusted by all other extremists (Fig. 3A).

But what happens if the charismatic leader dies or is otherwise removed from this constellation?[7] As can be seen from Fig. 3B, the opinions of the extremists nevertheless remain stable. The extremist group survives the elimination of its charismatic leader and preserves its internal stability. In fact, P2 moves up in the internal hierarchy and constitutes the new exclusive authority in the group (Fig. 3A). What distinguishes P2 as a potential successor is the fact that she only trusted the former leader and nobody else in the group, whereas the other group members already before the "death" of the former leader invested trust in P2 and selected her in this way as "crown prince". The successor was already in place.

As in the previous section we again slightly vary the experimental design to explore variations in the significance of external trust relations. In this experiment the members of the extremist group P3 and P4 do not only trust the other extremists P2 and P5, but also trust the "loner" P1 (Fig. 4A, $t = 0$). Consequently, the opinions of P3 und P4 (0.8) in the initial equilibrium are positioned between the opinions of the charismatic leader (0.9) and the "loner" (0.6). This constellation is also endogenously stable, but in this case the extremist group dissolves as soon as the charismatic leader P5 is removed (Fig. 4B). Without P5 the balance between extremist and external authorities is changed from the point of view of P3 and P4. After the disappearance of their highly trusted leader, P3 and P4 at first tent towards the "loner" P1. But in adjusting their first-order opinions to the

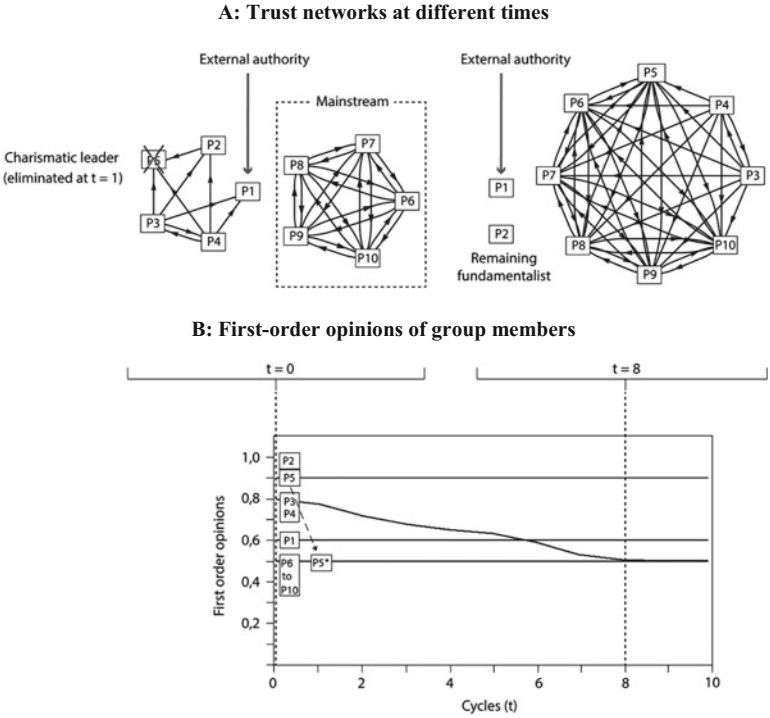


Numeric specification in Appendix 2.

Fig. 3 Simulation of the stability of extremist groups, sufficient exclusivity of trust relations

opinions of P1, eventually also the opinions of the mainstream inhabitants were included in their confidence interval. As the trusted persons with moderate opinions greatly outnumber the trusted persons with extremist opinions, P3 and P4 depart more and more from the remaining extremist until they do not trust her at all and are integrated fully in the mainstream.

Because the initially stable extremist group has failed to cut all external trust relations, it collapses when the charismatic leader is eliminated. This outcome of the simulation suggests that not only for the emergence but also for the maintenance of extremist groups it is crucial that they establish and



Numeric specification in Appendix 2.

Fig. 4 Simulation of the stability of extremist groups, insufficient exclusivity of trust relations

preserve particularistic in-group trust. In the long run, only such extremist groups can survive which successfully prevent external trust relations of their members and are ready to undertake serious efforts in providing resources to secure epistemic seclusion, social isolation and their "crippled epistemology".

Hypotheses

The first results of our simulation model demonstrate that even with this simple prototype informative and interesting hypotheses about the conditions for the emergence and continued existence of extremist groups can be

generated. The simulations support and reproduce the empirical finding that charismatic leaders can play an essential role in the dissemination and stabilization of extremist world views. Furthermore, we can on the basis of this model simulate and understand the basic social mechanism through which certain persons are first established as leaders in a group. Lastly, the instability and the erosion of extremist groups could be explained as a result of opinion dynamics under modified conditions.

The core of the modeling is the mutual adaptation of first- and second-order beliefs, or, to put it more generally: the role of epistemic trust in the formation of beliefs. Only if one systematically considers beliefs which refer to concrete spheres of life as well as beliefs which deal with epistemic competence and trustworthiness can one accomplish a sufficient level of complexity to comprehend the origin, establishment, and erosion of epistemic authority and its possible influence on the conversion from moderate to radical and extremist convictions.

The proposed model is intended as a model for a general social-doxastic mechanism which underlies not only the epistemic dynamics in extremist groups but processes of opinion formation in other contexts as well. It can be applied, therefore, to majority opinions and mainstream convictions about religious or political issues as well as fashion trends, youth subcultures, or esoteric circles. From our point of view, it is not a variation of the basic mechanism of opinion dynamics that is decisive but the contextual conditions in which it operates.

In the case of extremism, we can derive the following hypotheses from our experimental simulations:

1. Trust in a potential ideological leader must not initially be especially strong or exclusive. Existing trust relations toward moderate persons could be eroded in the process of opinion formation. Not all members of an extremist group must therefore be social outcasts from the start.
2. Charismatic leaders can come from outside with only weak trust relations to members of a group at the outset. It can be sufficient for them to become a group leader if only single members of the group trust them. This allows for promising infiltration strategies which are targeted only at a few people.
3. Unshakable self-confidence combined with a general disregard for the competence of other persons is a crucial precondition to become a charismatic leader. Persons with lower self-confidence will tend to subordinate themselves more and more to such leader personalities.

4. Extremist opinions can gradually become plausible and must not be inside the opinion space of the majority of group members from the beginning. There can be a self-reinforcing process of radicalization which takes place stepwise and sequentially.
5. Weak trust relations with the mainstream can immunize a group against extremist opinions. Relatively small shifts in these relations can tip a development and a critical threshold can easily be exceeded. Therefore it is an important strategy of extremist groups to combat this hazard potential and sever their member's external trust relations by all means.
6. Weak trust relations with outsiders can undermine extremist opinions in a group. Persons who are not part of the mainstream but do not express a radical position can build bridges for reintegration of extremists into the mainstream.

These hypotheses can be put in a nutshell: *taking the opinions of others seriously can be sufficient to become an extremist!*

As already emphasized, simulation models are highly idealized reproductions of reality which cannot substitute empirical validation of theories and deliver explanations per se. However, the simulation experiments with our prototype elucidate that such models can have a significant heuristic value and are suitable to analyze the basic mechanisms of complex social dynamics and to generate fruitful hypotheses. In our case the results are an additional support for Russell Hardin's ingenious theory of the "crippled epistemology" of extremist groups, and we recommend it as an excellent framework for future research in this troubling field.

Appendix 1: The Model

The model describes the collective opinion formation in a group of n persons ($G = \{1, \ldots, n\}$) in discrete time steps ($t = 0, 1, \ldots$). Each person $i \in G$ at a given point of time t has precisely one first-order opinion, $x_i(t) \in$ (with $j = 1, \ldots, n$), and n second-order opinions, $y_{\{i,j\}}(t) \in [0.1]$ (with $j = 1, \ldots, n$). The set $V_i(t)$ of all group members who are trusted by person i at time t is defined as:

$$V_i(t) := \{j \in G : (|x_i(t) - x_j(t)| \leq \epsilon) \wedge (y_{i,j}(t) \geq y_{i,i}(t))\}. \quad (1)$$

whereby $\varepsilon \in \mathbb{R}$ is a confidence parameter. The first- and second-order opinions of a person i are modified according to the following dynamic rules

$$x_i(t+1) = \frac{1}{|V_i(t)|} \sum_{k \in V_i(t)} x_k(t) \quad (2)$$

$$y_{i,j}(t+1) = \frac{1}{|V_i(t)|} \sum_{k \in V_i(t)} y_{k,j}(t). \quad (3)$$

Appendix 2: Numeric Specification of the Simulation Experiments

A simulation experiment is numerically completely specified by

- the group size n,
- the confidence interval ε,
- the initial first-order opinions

$$\mathbf{X}(0) = \begin{pmatrix} x_1(0) \\ \vdots \\ x_n(0) \end{pmatrix},$$

the initial second order opinions

$$\mathbf{Y}(0) = \begin{pmatrix} y_{1,1}(0) & \cdots & y_{1,n}(0) \\ \vdots & \ddots & \vdots \\ y_{n,1}(0) & \cdots & y_{n,n}(0) \end{pmatrix}.$$

Emergence of extremist groups (Fig. 1)

- Group size: $n = 10$,
- Confidence interval: $\varepsilon = 0.33$,
- Initial first- and second-order opinions:

$$\mathbf{X}(0) = \begin{pmatrix} 0.9 \\ 0.6 \\ 0.55 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix},$$

$$\mathbf{Y}(0) = \begin{pmatrix} 0.9 & 0.5 & 0.5 & 0.5 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.9 & 0.5 & 0.5 & 0.5 & 0.5 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.9 & 0.5 & 0.5 & 0.5 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.9 & 0.5 & 0.5 & 0.5 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{pmatrix}.$$

Emergence of extremist groups (Fig. 2)

- Group size: $n = 10$,
- Confidence interval: $\varepsilon = 0.33$,
- Initial first- and second-order opinions:

$$\mathbf{X}(0) = \begin{pmatrix} 0.9 \\ 0.6 \\ 0.55 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix},$$

$$\mathbf{Y}(0) = \begin{pmatrix} 0.9 & 0.5 & 0.5 & 0.5 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.9 & 0.5 & 0.5 & 0.5 & 0.5 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.9 & 0.5 & 0.5 & 0.5 & 0.5 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.9 & 0.5 & 0.5 & 0.5 & 0.5 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{pmatrix}.$$

Stability of extremist groups, endogenous stable situation with charismatic leader (Fig. 3, $t = 0$)

- Group size: $n = 10$,
- Confidence interval: $\varepsilon = 0.25$,
- Initial first- and second-order opinions:

$$\mathbf{X}(0) = \begin{pmatrix} 0.6 \\ 0.9 \\ 0.9 \\ 0.9 \\ 0.9 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix},$$

$$\mathbf{Y}(0) = \begin{pmatrix} 0.7 & 0.05 & 0.6 & 0.6 & 0.6 & 0.65 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.3 & 0.8 & 0.45 & 0.45 & 0.9 & 0.2 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.55 & 0.5 & 0.5 & 0.8 & 0.35 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.3 & 0.55 & 0.5 & 0.5 & 0.8 & 0.35 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.3 & 0.8 & 0.45 & 0.45 & 0.9 & 0.2 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{pmatrix}.$$

Stability of extremist groups, without charismatic leader (Fig. 3, $t>0$)

- Group size: $n = 10$,
- Confidence interval: $\varepsilon = 0.25$,
- Initial first- and second-order opinions:

$$\mathbf{X}(0) = \begin{pmatrix} 0.6 \\ 0.9 \\ 0.9 \\ 0.9 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix},$$

$$\mathbf{Y}(0) = \begin{pmatrix} 0.7 & 0.05 & 0.6 & 0.6 & 0.6 & 0.65 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.3 & 0.8 & 0.45 & 0.45 & 0.3 & 0.2 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.55 & 0.5 & 0.5 & 0.4 & 0.35 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.3 & 0.55 & 0.5 & 0.5 & 0.4 & 0.35 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{pmatrix}$$

Stability of extremist groups, endogenous stable situation with charismatic leader (Fig. 4, $t = 0$)

- Group size: $n = 10$,
- Confidence interval: $\varepsilon = 0.25$,
- Initial first- and second-order opinions:

$$\mathbf{X}(0) = \begin{pmatrix} 0.6 \\ 0.9 \\ 0.8 \\ 0.8 \\ 0.9 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix},$$

$$\mathbf{Y}(0) = \begin{pmatrix} 0.7 & 0.05 & 0.6 & 0.6 & 0.6 & 0.65 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.7 & 0.8 & 0.45 & 0.45 & 0.9 & 0.2 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.7 & 0.55 & 0.5 & 0.5 & 0.8 & 0.35 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.7 & 0.55 & 0.5 & 0.5 & 0.8 & 0.35 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.7 & 0.8 & 0.45 & 0.45 & 0.9 & 0.2 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{pmatrix}.$$

Stability of extremist groups, without charismatic leader (Fig. 4, $t > 0$)

- Group size: $n = 10$
- Confidence interval: $\varepsilon = 0.25$
- Initial first- and second-order opinions:

$$\mathbf{X}(0) = \begin{pmatrix} 0.6 \\ 0.9 \\ 0.8 \\ 0.8 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix},$$

$$\mathbf{Y}(0) = \begin{pmatrix} 0.7 & 0.05 & 0.6 & 0.6 & 0.6 & 0.65 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.7 & 0.8 & 0.45 & 0.45 & 0.3 & 0.2 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.7 & 0.55 & 0.5 & 0.5 & 0.4 & 0.35 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.7 & 0.55 & 0.5 & 0.5 & 0.4 & 0.35 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{pmatrix}.$$

Notes

1. Russell Hardin does not like the term "trust" in this context because he wants to reserve the use of this term for relations with "strong" ties (Hardin 2009, 26). Insofar we use a thin concept of trust which also includes relations which are impersonal but share important aspects with personal trust relations such as dependence or risk-taking. However, this is a terminological, not a substantial point of departure (cf. Baurmann 2010b).
2. The explanatory significance of such modeling is discussed in a special issue of *Erkenntnis* (vol. 70, no. 1, January 2009) "Economic Models as Credible Worlds or as Isolating Tools?" with contributions among others by Nancy Cartwright, Till Grüne-Yanoff, Tarja Knuuttila and Robert Sugden.
3. For an application of this model to a "veritistic" issue cf. Betz et al. (2013).
4. In this prototype, we do not differentiate between the attribution of competence and veracity but subsume both under "competence".

5. But in contrast to the Lehrer-Wagner model, the competence degrees are not used in our model as weights for averaging but only to select trustworthy persons.
6. Deffuant et al. (2002) and Deffuant (2006) refine the bounded-confidence model to study the dynamics of polarization and radicalization processes but they do not consider second-order opinions. The same applies to a recent publication by Hegselmann and Krause (2015) in which they explicitly deal with the dissemination of extremist beliefs but without including the formation of epistemic trust relations.
7. Technically the charismatic leader P5 is not removed from the simulation but becomes part of the mainstream.

References

Baurmann, M. 2007. Rational Fundamentalism? An Explanatory Model of Fundamentalist Beliefs. *Episteme: Journal of Social Epistemology* 4: 150–166.

———. 2010a. Fundamentalism and Epistemic Authority. In *Democracy and Fundamentalism*, The Tampere Club Series, ed. A. Aarnio, vol. 3, 71–86. Tampere: Tampere University Press.

———. 2010b. Kollektives Wissen und epistemisches Vertrauen. Der Ansatz der Sozialen Erkenntnistheorie. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*. 50 (Sonderheft): 185–201.

Baurmann, M., G. Betz, and R. Cramm. 2014. Meinungsdynamiken in fundamentalistischen Gruppen. Erklärungshypothesen auf der Basis von Simulationsmodellen. *Analyse & Kritik* 36: 61–102.

Betz, G. 2006. *Prediction or Prophecy? The Boundaries of Economic Foreknowledge and Their Socio-Political Consequences*. Wiesbaden: Deutscher Universitats-Verlag.

———. 2010. *Theorie dialektischer Strukturen*. Frankfurt am Main: Vittorio Klostermann Verlag.

Betz, G., M. Baurmann, and R. Cramm. 2013. Is Epistemic Trust of Veritistic Value? *Etica & Politica – Ethics & Politics* XV: 25–41.

Deffuant, G. 2006. Comparing Extremism Propagation Patterns in Continuous Opinion Models. *Journal of Artificial Societies and Social Simulation* 9 (3).

Deffuant G., F. Amblard, G. Weisbuch, and T. Faure. 2002. How Can Extremism Prevail? A Study Based on the Relative Agreement Interaction Model. *Journal of Artificial Societies and Social Simulation* 5 (4).

Hardin, R. 1992. The Street-Level Epistemology of Trust. *Analyse & Kritik* 14: 152–176.

———. 2002. The Crippled Epistemology of Extremism. In *Political Extremism and Rationality*, ed. A. Breton et al., 3–22. Cambridge: Cambridge University Press.

———. 2009. *How Do You Know?: The Economics of Ordinary Knowledge*. Princeton: Princeton University Press.

Hedström, P., and R. Swedberg, eds. 1998. *Social Mechanisms*. Cambridge: Cambridge University Press.

Hedström, P., and P. Ylikoski. 2010. Causal Mechanisms in the Social Sciences. *The Annual Review of Sociology* 36: 49–67.

Hegselmann, R. 2004. Opinion Dynamics – Insights by Radically Simplifying Models. In *Laws and Models in Science*, ed. D. Gillies, 19–46. London: King's College Publications.

Hegselmann, R., and A. Flache. 1998. Understanding Complex Social Dynamics: A Plea for Cellular Automata Based Modelling. *Journal of Artificial Societies and Social Simulation* 1 (3).

Hegselmann, R., and U. Krause. 2002. Opinion Dynamics and Bounded Confidence: Models, Analysis and Simulation. *Journal of Artificial Societies and Social Simulation* 5(3).

———. 2006. Truth and Cognitive Division of Labour First Steps Towards a Computer Aided Social Epistemology. *Journal of Artificial Societies and Social Simulation* 3 (9).

———. 2015. Opinion Dynamics Under the Influence of Radical Groups, Charismatic Leaders, and Other Constant Signals: A Simple Unifying Model. *Networks and Heterogenous Media* 10: 477–509.

Lehrer, K., and C. Wagner. 1981. *Rational Consensus in Science and Society: A Philosophical and Mathematical Study*. Dordrecht: D. Reidel.

Violence and Politics in Northern Ireland: IRA/Sinn Fein's Strategy and the 2005 Disarmament

Carolina Curvale

Introduction

Ethnic conflicts have become increasingly common in the world we live in, some of which ultimately result in bloodshed. Russell Hardin's *One for All: The Logic of Group Conflict*, published in 1995, offered an alternative view to the prevailing explanations of ethnic conflict as the result of emotional behavior. Hardin shook up the field by proposing that group identification, conflict, and violence could be understood from rational choice theory. This chapter seeks to provide an account of the Northern Ireland case from this perspective.

A large body of literature has been devoted to explaining the Northern Ireland case. Nevertheless, there is no consensus regarding the definition of the problem itself (McGarry and O'Leary 1995). Several cleavages[1] divide Northern Ireland's population into two ethnic-religious groups; even

I thank Kim Stanton for valuable comments. All errors remain mine.

C. Curvale (✉)
Department of Political Studies, FLACSO Ecuador, Quito, Ecuador

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_11

though religion is the key ethnic marker of the groups, hostilities go beyond national and religious identification, including asymmetries in the distribution of economic and political power. For decades Protestant and Catholic paramilitary organizations fought over non-reconcilable claims on how to define the boundaries and political status of Northern Ireland, a struggle responsible for approximately 3500 deaths.[2] While Unionists (predominantly Protestant) prefer to remain a part of the United Kingdom, Nationalists/Republicans (predominantly Catholic) advocate for a unified Ireland, whose territory would cover the entire island.

Why did violence last? In the late 1960s, the Northern Ireland Civil Rights movement organized several campaigns to demand fair treatment of the Catholic community. Catholics were brutally repressed by the Protestant official police and suffered physical aggressions perpetrated by extremist Protestant groups. I argue that this situation offered an opportunity for the Irish Republican Army (IRA) and the Sinn Fein (SF)—Provisional IRA's political wing—leaders to gain access to power by sustaining violence and escalating the conflict through paramilitary activity. While much progress has been made—in fact the number of conflict-related deaths has declined significantly over the last decade—the peace-building process is still ongoing (Power 2011).[3]

IRA/SF failed to gain support among the Catholic population through its methods of terror. In the 1999 elections to the European Parliament, it obtained only 17% of the votes and 22% in the 2001 Westminster elections, which barely accounts for half of the Catholic population. But prospects for peace led to electoral improvements: In the 2003 Assembly election, SF obtained 23.5% of votes and 26.3% in the June 2004 European Parliament election. Before the peace process began, popular support for SF was even lower.[4] Estimates indicate that only between 500 and 600 people were active IRA participants (Taylor 1999b: 363).[5] This suggests that violence was not massively supported and, in turn, that ethnic identifications had led to sustained violence only through the action of the most extreme minded individuals in each group.[6] Loyalist paramilitary organizations did not face a better outlook. For a total population of about 1,600,000 people,[7] the number of activists perpetrating violent acts on both sides was almost insignificant.

This chapter proceeds as follows. In section "Group Identification," group identification of Catholics and Protestants is discussed departing from reviewing the cleavages that divide the two groups. Section "IRA's Membership" analyzes the factors that allowed the IRA to survive as an

organization and how extremist militants decide to join the organization. Section "SF's Political Strategy" focuses on SF's strategy in its quest to gain political power. Section "Concluding Remarks" concludes.

Group Identification

In Hardin's account, there is a clear difference between identification and mobilization on behalf of the group identification of choice. Identification is one possible equilibrium of a simple strategic problem of coordination. It does not even pose the demands of a collective action problem, since we need not provide additional incentives to promote participation or penalize defectors. We maximize our happiness by behaving similarly in some general aspects of our lives, from enjoying the same drinks to sharing a worldview. Surrounding ourselves with people who share our preferences, all the more so when it includes important topics such as religion, reinforces our identification. When the shared preferences include distributional considerations with other groups, as in the Northern Ireland case, group membership offers additional benefits, like the prospect to control resources via public office. When we combine the concept of identification as the product of coordination with the notion of the "epistemological comforts of home," the plausibility that individuals choose in a rational manner (that is, following self-interested motivations) a particular group loyalty increases dramatically.

Several factors shape group identification of the Northern Irish, resulting in the division of the population into two groups that are generally referred to by their religion markers: Catholics and Protestants. One way to measure the extent of this division is to look at election results. Unionists (Protestants) had historically received about 60% of electoral support, while Nationalist (Catholics) parties keep the remaining 40%, which in turn reveals the status of the latter as an important minority. But even when the conflicts go beyond party identification, the results obtained through elections do not fail to capture the depth of division between the two subnational groups. As Rose suggested, the cleavages dividing the population tend to be self-reinforcing, meaning that they marshal almost the same sets of individuals across different definitions of belonging (Rose 1971). However, a main component of ethnic identification, language, unites rather than separates Catholics and Protestants. According to the 2011 census only 10.7% of the population has some hability in Gaelic (English

being the predominant language) and only 0.24% uses Gaelic as the main language at home.[8]

The odd distribution of population in Northern Ireland is the result of partition, introduced by the Government of Ireland Act, issued in May 1920 by the British government. The issue of partition is key in IRA's motives for action. The IRA (by then called Irish Volunteers) performed its first act on stage in the Easter Rising in 1916, which took place in Dublin. It was conformed by about 200,000 people. The fact that England was involved in World War I, gave an opportunity to issue the proclamation declaring Ireland a Republic. The British repression was very intense, and had the effect of fostering nationalist support in Ireland, through IRA and its political wing, Sinn Fein. Until 1918, Sinn Fein adopted a position of political abstention: when they won elections, they did not accept their seats in Westminster Parliament. According to O Heithir, "the majority of Irish people were still more willing to support political struggle than military rebellion" (1997: 12–3). A special event took place that had the effect of increasing nationalist ranks: the British government intended to link conscription during the war to home rule, which was unacceptable for many Irish. In the meantime, Unionists were also discontent with the Home Rule situation and in favor of a more direct relationship with England. The third political force, the Irish Party, was in favor of Home Rule under the Crown. In 1919, Sinn Fein declared Independence and the Irish Volunteers took the official name of Oglaigh na hEirean, the Gaelic expression for IRA. The Government of Ireland Act followed this procedure and was the result of the national struggle in the south of the island.

Partition established the creation of two governments—Northern and Southern—each responsible for the maintenance of peace.[9] As a result, a large minority of Catholics remained in Northern Ireland, but the reverse did not happen in the Irish Free State, where according to the 1926 census, only 5.5% of the population declared themselves to be Protestants. It should also be noted that in the 1911 census the percentage was almost 8, and was reduced after partition, which suggests that some Protestants migrated to Northern Ireland.[10] Since the Catholic minority was so significant, homogenizing the groups territorially would have involved huge transfers of population, which was not feasible; besides, by then the expectation to achieve a unified Ireland was a vivid hope among Catholics. Over the years, the size of the groups has progressively balanced: the 2011 census reported that 48% Protestants and 45% Catholics reported to have been brought up in that religion.

The reunification of Ireland is the goal of nationalist parties in Northern Ireland—moderate and ultra—although they differ in the means of its attainment.[11] Conversely, Unionists prefer to remain politically and culturally close to the United Kingdom and to maintain partition. Since both claims are irreconcilable, there are two competing nationalisms in Northern Ireland's conflict (Ruane and Todd 1998) fighting over the control of an indivisible unit.

The terms Unionist and Protestant appear almost interchangeably in the literature, as do the labels Catholic and Nationalist, which is not a completely accurate picture. The ancient root of the conflict implies that several generations were born, raised, and died in the soil of the island of Ireland. In 1968, 20% of Protestants identified themselves as Irish, while 15% of Catholics defined themselves as either British or Ulster (Rose 1971). A similar pattern can be found in the case of Protestants in the Republic of Ireland, who formerly identified with the Union. Over time, they began to self-identify as Irish Protestants (Whyte 1990). This information reinforces the religious explanation versus the British-Irish dichotomy.

Other sources of hostilities between the two groups include economic issues. By being a part of the United Kingdom, Northern Ireland receives a subsidy. McGarry and O'Leary nicely depict the reductionism involved in the interpretation of the conflict solely in terms of economic interest:

> While some Irish nationalists allege that Protestants are unionists for economic reasons, they would never concede that they themselves should fully integrate with the United Kingdom if this was in their economic interest. Unionist integrationists, by contrast, argue that Catholics should be happy with equality of opportunity and prosperity in the UK, but they themselves would never accept the same offer within a united Ireland. (1995: 306–7)[12]

Perhaps the most relevant aspect of the economic dimension of the conflict is that Catholics have generally been poorer than Protestants. The fact that Protestants were a majority in Northern Ireland gave them control over public policy and government structure. The control over state resources and its use against the minority raised several social and civic claims in the context of the Protestant Stormont regime, which ruled Northern Ireland from 1921 to 1972.

Addressing these allegations of discrimination, the Northern Ireland Civil Rights Association (NICRA), made up of unionists, nationalists, and workers, organized several peaceful demonstrations in Derry (O Heithir

1997; Rose 1971). Among the leadership of the civil rights movement were a few nationalists linked to the IRA, but also unionists and representatives of the labor force. Therefore, it cannot be assumed that the IRA coordinated NICRA's actions. McGarry and O'Leary (1995: 312) argue that most of the participants were nationalist; even though the movement was meant to be and was born as a civil association, initiated by professionals and without any particular partisan affiliation or ambition (Rose 1971: 101–103). As Bell notices "it would be fair to say that the Civil Rights movement had far more influence on the IRA than the reverse" (Bell 1997: 358).

The NICRA and People's Democracy (a student association) articulated demands of discrimination in housing and employment against Catholics, the abolition of repressive legislation, and political rights. Evidence has shown that while Catholics were indeed discriminated against housing policy in Protestant counties, the bias favored Catholics in areas controlled by Catholic councilors. Therefore, housing discrimination against Catholics was not systematic (Rose 1971: 293). The alleged discrimination in terms of employment presents a less ambiguous situation. Data gathered in 1971 show that only 2.7% of Catholic representation in the public and private sectors, which increased to 4.4% in 1991 (Gudgin 1999: 108). Under the Stormont parliament, it is estimated that 10,000 Catholic workers were fired and 23,000 Catholics were driven out of their homes (Taylor 1999a: 25).

NICRA also demanded the abolition of the Special Powers Act (SPA). Introduced in 1922, the SPA gave the Northern Ireland Minister for Home Affairs sweeping powers to fight subversion without much concern for civil liberties.[13] Other demands were the removal of property qualification in local elections (which tended to prevent Catholics from voting since they were less wealthy), and the end of the gerrymandered local government boundaries.

We have mentioned three important factors that may determine group identification -ethnicity, religion and economic and political interests.[14] Now, how did mobilization on behalf of group identification begin? How did violence start? The Royal Ulster Constabulary and the B-Specials, the State military forces, excessively repressed the civil rights demonstrations of 1968. Protestant extremist groups also attacked the participants, and further aggressions took place in Catholics and Protestant towns. At last, a peace line was made by British Army troops, and London ended home rule and put Northern Ireland under Britain's direct government. In Hardin's framework, these demonstrations were "the tipping phenomena" that

triggered violence. In addition, I argue that violence escalated as a consequence of the opportunistic behavior of the IRA and the SF leadership, which capitalized on the conflict between the two groups and the violence surrounding the NICRA protests. What began as a demand for equality would end up, through the timely action of the IRA, in a long-lasting war that reintroduced the question of partition as the main issue.

IRA's Membership

In 1968, at the beginnings of the troubles, public support for the use of violence was 51% in the Protestant community, and only 13% among Catholics. Data gathered in 1998 showed that 31% of Protestants expressed some level of sympathy for the loyalist paramilitaries, while 28% of Catholics sympathized with republican paramilitaries.[15] If the IRA strategy of violence had not been widely supported, how did it manage to maintain its structure and activity for the past 30 years? One factor that played an important role was the historical opportunity the IRA had to recruit members. At the peak of the civil rights movement, the Protestant State's discrimination against Catholics produced an intense feeling of resentment, particularly among young people. The majority of the participants were students or working-class unemployed people, some of them socialists, who might have been influenced by ideas of revolution—but of course, not all of them did join.

In addition to feelings of resentment, the state's violent response and further violence perpetrated by Protestant extremist groups generated an environment of fear and insecurity. Violent riots were increasingly taking place. Spontaneous community-defense committees were organized in the face of the aggressions. Just being a member of (or living in) a Catholic community was a reason for being attacked by Protestants. As Hardin notes,

> Self-defense against possible (not even actual) attacks suffices to motivate murderous conflict. Risk aversion is enough. And the risk, unfortunately, of not preemptively attacking may be heightened by the fact that the other side – such as an ethnic group – cannot commit to not attacking, and therefore cannot be trusted beyond what can be inferred from their interests. (1995: 143)

Northern Ireland in the late 1960s fits this description. Old fears and resentments were pushing things forward. It might have ended when the British Army troops made a peace line: they were providing order, but at the

same time the troops represented a reason for raising nationalist claims and further aggression. Once violence began, identification was reinforced by opposition to the other group. The mere fact of "being a Catholic" was a reason for being at risk, regardless of political affiliation. Mutual fears between the groups gave strength to paramilitary groups as part of a defense strategy. In the absence of state protection, the IRA filled the gap, by having the organizational skills and the know-how for engaging in armed struggle.

Even when the broader population did not directly get involved in the armed struggle and even opposed it, once the cycle of violence began it was presumably better for individuals to receive the benefits of protection involved with membership in the community, since the mere fact of belonging to the group was endangering. This could be thought of as a positive externality: a military force would monitor the town. Of course, that did not guarantee complete security, but at least provided some probability of being protected. This reasoning could have been suitable at the start of the conflicts, as a defense strategy. Once violence escalated and the groups adopted the logic of retaliation, no one was better off. Internal organizational dilemmas led to several killings by the IRA of members of the Catholic community. Between 1969 and 1998, the IRA killed 99 Catholic civilians as "unintended targets," while 23 more died in the hands of other republican paramilitary groups that had separated from the IRA.[16]

Why would an individual join the IRA? Responding this question involves a possibly insurmountable methodological problem, which is entering the minds of extremists and trusting that we get a truthful answer. But we may get some insights from other pieces of information. Even though many of the individuals that joined the IRA were unemployed, the IRA stipend was so low (Coogan 1993) that it was an unlikely incentive. Nor were their motives related to some kind of psychological distortion. As Heskin recounts,

> In regard to the assertion that terrorist groups contain strong psychopathic elements, the argument here tends to be speculative and circular. It is speculative in so far as, to my knowledge, there is no psychological evidence that those who have been involved in terrorist activities are, in fact, diagnosable psychopathic or otherwise clinically disturbed. Indeed, what little evidence there is of this type points in exactly the opposite direction. (1980: 78)

Field research confirms this assertion. Through personal interviews, White has studied the motivations for IRA enrollment (White 1989). The

results of his work show that those who join the IRA feel a social commitment to their community, have experienced state repression, and believe that organized political violence will produce social change[17]; it was also noticed that injustice impinges on their national identity. White concludes,

> The data show that the decision to become involved in political violence is influenced by state repression and interaction with other people experiencing this repression. The data also show that this decision is an emotional, political and rational one. (1989: 1295)[18]

Since a number of IRA recruits were unemployed and had experienced state violence, we could think that their perceived set of options was more restricted than that of other Catholics (i.e. professionals). Individual's rational decisions are to a great extent a product of information, social experiences, and beliefs, and they constitute a constraint on the set of choices perceived as available. As Hardin asserts, "[...] what it is rational (in one's interests) to do depends on who one is in the sense that it depends on what knowledge one has" (1995: 17). Yet many working-class individuals and unemployed Catholics supported the SDLP, the constitutional nationalist party. In light of White's analysis, we may think that they did not experience direct state repression; but certainly many of them might have known someone very close to them that did. It is estimated that "Catholics are twice as likely to have been intimidated when compared to Protestants, and they are about one-third more likely to have been the victim of a violent incident" (Hayes and McAllister 2001).

Joining the IRA imposes high costs on individuals: a militant risks losing her life in the pursuit of group benefits. Hardin points out that extremists' judgment of their perception of life options and value may be blurred by epistemological ignorance, that is, "the suppression of ordinary understandings" (1995: 164–5) that distorts the cost-benefit calculation of participation. This may well be a result of a successful indoctrination process. A 2009 incident may serve as an example of how epistemological distortion could operate. On March 7, 2009, two British soldiers were shot dead in Atrim as they were getting a pizza delivery, and the delivery workers were injured. The Real IRA claimed the incident and justified the pizza workers' injures as follows: "In delivering the pizza, they were serving the British state occupation of Ireland" (quoted in Sanders 2011: 238).

Once an individual expressed his/her will to join the organization, though in a secret manner,[19] an answer could take several months. After

that, they were "actively discouraged from joining with warnings of the fate that would probably await them, prison or death" (Taylor 1999b: 89). Exit from the organization was not punished, since willingness to stay was necessary in order to prevent betrayal; but having been a member of the IRA made a former recruit a potential target of loyalist paramilitary violence. The IRA appealed to tradition, culture and the Gaelic language in order to justify the rightness of the republican cause. Every new recruit was given a "Green Book," where the rules of conduct where established with topics ranging from the political goals of the movement to the expected private behavior of a volunteer. The following extract shows the intensity of the volunteer's compromise demanded by the IRA, when dealing with the situation of eventual interrogation:

> The best protection while being interrogated is LOYALTY to the Movement. This implies LOYALTY to all YOUR COMRADES and PROTECTION of all members of the Movement. Again commitment to the aims and objectives of the Movement, a deep and unmoving POLITICAL COMMITMENT to the ideas of the Socialist Republic, CONSTANT AWARNESS that you are a REVOLUTIONARY with a sound POLITICAL base, NOBLE and JUSTIFIABLE CAUSE, and deep and firm belief that those holding you and interrogating you are MORALLY WRONG, that you are SUPERIOR in all respects, because your cause is RIGHT and JUSTIFIED. (Coogan 1993: 430)[20]

This fragment also depicts the extent of the definition of "us" and "the others." "We are right," "they are wrong." Hunger strikers are an extreme example of the extent of the compromise of an IRA volunteer with the republican cause. The IRA presumably managed to distort their recruit's perception of what was really going on—people were dying, and the struggle did not stop and was unable to lead to a solution—in order to pursue and justify the use of violence. Every loyalist member was a suitable target and sometimes, a civilian "Protestant," even though not politically active, was killed. The most serious consequence of this way of thinking is that it narrows to such extent the individual's ability to assess the situation that they cannot conceive alternative courses of action. Over the years, some IRA members eventually left the organization, some died, but still today there are a few intense believers.

Another factor that contributed to the IRA's survival was its mode of operation. The IRA functioned with a cell structure that allowed it to be less

susceptible to the control of the official forces and therefore protected the organization. On the flip side, the cell structure made lack of coordination within the organization more likely, and therefore over time multiple internal splits occurred—some of them out of disagreements on strategy, others due to internal struggles for power. Republican terrorists are divided into two main groups: the IRA, which is joined by the Irish National Liberation Army (INLA) in its decision of cessation of fire, and the dissident Republicans. The latter is divided into two organizations, namely, the Real IRA and the Continuity IRA. Both opposed the Good Friday Agreement and returned to violence.[21] A similar pattern of internal splits governs loyalists paramilitary organizations.

SF's Political Strategy

In this section, I argue that SF and the IRA behaved strategically in order to maximize its chances to access power via the use and the threat of the use of violence, especially since 1982 when SF started to participate in elections as a party in its own right. Before that year, independent nationalist candidates and other groups captured the votes that were later transferred to SF; in fact, SF itself supported some of these candidates. During the 1973–1982 period, the number of votes obtained by these groups was no greater than 16%[22] (this figure includes the candidates for the Civil Rights movement in 1973, which SF supported).

The start of the troubles was a key opportunity for the IRA to gain relevance as a major player in the political game, by providing defense against the Protestant State's aggressions. The IRA leadership requested support from Dublin, but it was denied, triggering a split within the IRA into Provisionals and Officials.[23] The Provisional IRA[24] prevailed after a couple of years of competition. In turn, SF had to compete for support with the SDLP, the nonviolent republican alternative.

A major incident gave broader support to the IRA as the "private police." In January 1972, 13 civilians were shot dead by the Parachute Regime in Derry, an event referred to as "Bloody Sunday." In reprisal, the IRA killed seven people in a Parachute Brigade town. The retaliatory logic of the paramilitary groups reinforced violent demonstrations and raised conflict-related deaths. Demands for protection transformed into active attacks against "the other group." As a result of the extended violence, the British Government put Northern Ireland under Westminster direct rule in March 1972. This marked the end of the Protestant Stormont regime.

The year 1972 constituted the peak of paramilitary mobilization and violence, registering the highest number of deaths (476) in the history of the conflict, of which 55% were attributed to republican activists, while loyalists accounted for 23.5%.[25] The goal of the IRA was not to protect the Catholic community, but to force Britain to withdraw. The talks held that year between the Provisional IRA leadership and the British government did not get anywhere. The empowered Provisional IRA planned to continue to escalate the conflict until the British government conceded their demands.[26]

The IRA was further invigorated by the hunger strikes. In 1981, ten Republican hunger strikers who demanded to be treated as political prisoners, died in prison. Both the Provisional IRA and the British government maintained a very strong position.[27] As the IRA gained greater support, it was less likely to soften its position, thus engaging in a zero sum game. In the 1981 election, Bobby Sands, an IRA hunger striker, won 10.8% of the votes, but only through the banner "H-Block," which was the name of the section where he was in jail. The hunger strikes fortified the nationalist discourse by demonizing Britain: young men were dying for the "republican cause." At the same time, the permanent exposition to violence and the paramilitary guerrilla reinforced the IRA's position as a community protecting force.

Presumably encouraged by the triumph of the hunger strikers in 1981, SF decided to change its strategy and participate on its own right in elections. This was intended to be a new way to involve people in the movement, who were previously passive supporters: "Not everyone can plant a bomb, but everyone can plant a vote" (Irvin 1999: 91). However, this was not as promising as expected: in the 1982 election, SF's performance was meager. The share of the vote in comparison with the "H-Block" banner in the previous election fell by over 50%, which was not enough for winning a seat, suggesting that the electoral impact of the hunger strikes was not transferable to SF. This led the party to decide in favor of not contesting the following elections, which took place on November that same year (Coogan 1993: 381).

In 1983 (Westminster elections) and 1984 (European elections) SF got about 13% of the votes, while in the 1985 local government elections its vote share dropped by 1%. This decline in votes took place the same year when negotiations for reaching peace between the Irish Republic and the British Government were advancing, without the participation of SF.[28] Until 1986, SF supported a policy of abstentionism. Taylor notes that

Gerry Adam's (SF's leader) argument for changing that policy was that "the principle was no longer relevant in the latter part of the twentieth century"[29] (Taylor 1999b: 30). This move could be interpreted as recognition that remaining outside of the negotiation table would not be a successful strategy. Only 11% of the votes were obtained in the 1987 Westminster election. In 1988, the IRA failed a key operation in Gibraltar along with other tactical mistakes (Coogan 1993). In addition to the declining political support, the spirit in the IRA's ranks demoralized. In this context, Adams met Hume (SDLP's leader) in January 1988, but no joint action could be taken if the IRA continued its terror campaign.

SF received only 9% of the votes in the 1989 elections to the European Parliament (June 1989): it was its worst electoral performance ever. After many years of fighting, there were still no signs of victory. The IRA/SF was still far from reaching its goal of reunification and of being close to the management of state power. Since the beginning of the troubles, the Catholic community and the IRA were pursuing essentially different things. While the former demanded equal treatment and respect, the latter essentially demanded unification, leaving no room for negotiations.

In April 1992, Adams lost his Westminster seat in West Belfast, which was a serious warning. The original hard-line strategy had yet to change more its position in favor of an agreed solution. Adams appealed to Hume once more in order to negotiate in April 1993, this time willing to make concessions. They secretly signed an agreement in September 1993, which finally resulted in the 1993 Joint Declaration of Peace, signed by John Major, the British Prime Minister, and Albert Reynolds, the Irish Prime Minister. The document asserted that the people of Northern Ireland would decide its future as a political entity and addressed a demand for an IRA ceasefire. After some misunderstandings, the IRA finally announced its ceasefire in late August 1994. Six weeks later, the loyalist paramilitary organizations did the same. Therefore, the IRA was represented at the negotiating table and some steps were made with regard to the definition of the conditions for the decommissioning of arms. The IRA's cessation of fire took place soon after SF's worst electoral performance: the results of the 1994 elections to the European Parliament (June 9, 1994) were 9%, as in 1989. The move in the direction of self-binding the use of violence was politically profitable, since after that, votes in favor of SF rose. The IRA/SF position toward negotiations responded to the expected probability of success in the acquisition of power.

By February 1996, negotiations had reached a stalemate and the IRA once again returned to violence,[30] and was contested by loyalist paramilitary organizations. In the 1996 forum elections, held in June, SF got 15% of the votes. The British government tried to reinitiate negotiations, without success.[31] Finally, soon after the election of Mr. Blair in Britain, the IRA announced its definitive cessation of fire, which meant a passport to join the peace talks. After that, SF's electoral performance improved, reaching 17,65% of the vote in the 1998 elections to the Northern Ireland Assembly held in June. In April of that year, the British and the Irish Governments, and all Northern Ireland's political parties signed the Good Friday Agreement.[32] This document, which is considered consociational in Horowitz's sense, contains several crucial initiatives for reaching peace. Among them are the acceptance of the principle of consent for deciding the future of Northern Ireland, measures toward equality and respect of human rights, the reformation of the RU, the creation of cooperative institutions between the North and the South, recognition of self-identification as British or Irish or both, and the decommissioning of terrorist weapons. SF needed a new political strategy and its participation in the peace process was not possible without decommissioning of arms: "[...] the GFA would not have been possible without including Sinn Fein. Arguably, without the inclusion of republicans in the peace process through the 1990s, and ultimately in the executive, the IRA would not have been incentivized to (albeit slowly) decommission their weapons and may even have continued their campaign of violence" (McEvoy 2015: 82). Although the Agreement included some measures to address inter-group inequality, the IRA refused for 3 years after it was signed to proceed with the decommissioning of arms. This is consistent with a strategic use of the possibility to resort to violence, at least while electoral support strengthens. As Dimitrijevic (2001) points out, "[...] a terrorist act, regardless of the shock it produces, does not show that its authors have many followers: it may as easily demonstrate that there are too few adherents to support the cause by democratic, majoritarian means."

The power sharing institutions included in the Good Friday Agreement started to timidly function, but uncertainty prevailed as the peace process remained at a stalemate with regard to the IRA's decommissioning of arms. Of particular importance was the inclusion of the d'Hont system to allocate ministerial seats on the bases of parties' assembly seat shares, which provided assurances to SF that it would be a part of the governing coalition provided its strength in the assembly (McEvoy 2015: 68–69).[33] The agreement was subsequently endorsed in referendums in Northern Ireland and the

Republic of Ireland. In the 1998 assembly elections, SF obtained 18 out of 108 seats, SDLP got 24 and the unionist were narrowly divided with 58 seats. However, it took an additional 9 years until the Good Friday Agreement was finally implemented in 2007, when DUP and SF agreed to work together in government.

The 1999 direct European elections did not change much SF's electoral support.[34] SDLP managed to widen the electoral gap that separated it from SF, accounting for 6% more votes than in the previous election, although this gain was not obtained at the expense of SF. In this election the Unionist parties (Ulster Unionist Party (UUP) and Democratic Unionist Party (DUP)) and the Republican parties (SDLP and SF) received more votes at the expense of smaller parties. This might be an indicator of the fact that some voters changed their vote from small parties to the major ones that had a more relevant role in the peace process.

By February 2001, the lack of progress in disarmament led Trimble, the leader of the largest party in Northern Ireland at the time (the UUP), to lobby in London for a revision of the Good Friday Agreement. Although the loyalist/unionist paramilitary groups still kept their guns, Trimble claimed that SF should not be sharing legitimate power until it took concrete steps to get rid of its weapons. On the other side, SF sustained that decommissioning included the removal of the British army's guns as well. During this period both republican and loyalist paramilitary groups—allegedly the dissident—continued to perpetrate acts of violence.

The electoral results of the 2001 Westminster elections[35] showed polarization of the political system between Nationalists and Unionists, with the most hard-line groups in each side increasing their respective shares of votes. Indeed, SF accounted for 22% of the votes, reaching a historic peak, mainly due to the consolidation of its political control in the West.[36] It replaced SLDP as the largest nationalist party. On the unionist side, this general election also marked major gains for the hard-line DUP; but in this case the traditional major unionist party, the UUP, remained the largest party though intensively divided, but only one seat ahead of the DUP. Adding up the number of votes on each side, we still get a majority of Unionists, with 49% of the votes, while Nationalists gather 43% at the expense of other parties.[37] Both the 2003 Assembly election and the 2004 European Parliament election saw DUP and SF consolidate as the two largest pluralities.

Are these results the consequence of increased animosity between Catholics and Protestants? On the side of the Republicans, two things should be taken into account. First, SF's increased share of the votes can be interpreted

as the voters' wish to encourage the republican movement along its nonviolent path. Second, the electoral success of the SDLP (nonviolent alternative to SF) may have been hindered by the fact that the party had its own internal tensions, but the party cannot be said to have suffered a defeat. Its vote did not significantly drop; it retained all its Westminster seats and it lost only three of its councilors in the local elections.[38]

Since my concern is with the political strategy of the IRA/SF in its pursuit of power, we should look at SF's position toward political violence just before the election. One week before the 2001 general election, IRA issued a statement saying that it had held four meetings with the arms decommissioning body and that it had honored every statement it had made, but that the British government continued to renege on the issues of policing and demilitarization.[39] The message was clearly attempting to present IRA as willing to decommissioning arms only when "the others" complied with fair conditions. After the elections, no progress was made regarding decommissioning. First Minister of David Trimble -the UUP's leader- resigned, which accelerated the deterioration of the institutional arrangements adopted under the Good Friday Agreement. Both the British and the Irish Prime Ministers[40] proposed an Implementation Plan for the Good Friday Agreement, accompanied by a proposal for the IRA's decommissioning. The UUP rejected both, arguing that an actual action of decommissioning by the IRA was essential for the advancement of the talks; the UUP demanded the abolition by the British government of the Assembly and the institutions. In response, the IRA withdrew its proposal, and asked for fresh elections to end the deadlock. Let us recall that SF had gotten its highest electoral turnout in the immediate previous elections. According to polls, the "fresh elections" proposal was in line with the desire of 41% of voters.[41]

On August 10, 2001 the Secretary of State of Northern Ireland set a deadline of 6 weeks to solve the political crisis; he could decide to call for a review of the Good Friday Agreement, which would involve an indefinite suspension of the power-sharing government. Alternatively, he could opt for fresh Assembly elections. Three days before the deadline was reached, Gerry Adams asked the IRA to decommission. On October 23, the IRA made a historical shift in its position and begins to decommission by putting "beyond use" a significant amount of arms, which the Independent International Commission on Decommissioning witnessed. This event succeeded in containing the crisis and during early November Trimble (UUP) came back to power.

Why did SF offered to decommission in 2001 *after* it's the best election until that time? First, as was discussed above, violence had not proved profitable under the governing electoral dynamics. This IRA move had the additional advantage of highlighting in the eye of the public its willingness to cooperate toward peacekeeping as opposed to the loyalist street violence.[42] On the hole, SF appears to have realized that having the initiative of a nonviolent position conducive to peace could attain more gains in terms of political support.[43] Second, the virtual collapse of the 1998 Good Friday Agreement would have presented a worse scenario (going back to Westminster direct rule) than decommissioning. Finally, the costs of pursuing this strategy were not high. In truth the October 2001 decommission did not mean at all that IRA had lost its power.[44] Unionists are not convinced by this IRA move. In the 2002 Easter statement the IRA made no reference to further decommissioning, but did call on Irish nationalists to support Sinn Fein's political efforts.

The political process backtracked in October of 2002. The unionists threatened to quit the Assembly alleging suspicions of spying activity by the IRA, in response to which the British government assumed direct control. Negotiations resumed in March and April of 2003, but SF's ambiguity received a serious pledge for full disarmament by Prime Minister Tony Blair, who counted with widespread support, including the British media and the government of Ireland. The 2003 Assembly elections saw the rise of the DUP and SF as the largest pluralities, gaining each 30 and 24 seats respectively. The 2004 European Parliament elections showed similar results. However, two events damaged SF's public image: a bank robbery in December 2004 and a murder in January 2005 were linked to the IRA. In July 2005, the IRA made the historical announcement that it would unequivocally relinquish violence, although there have continued to be some violent incidents after that.

The IRA/SF's political strategy adjusted and oscillated as a function of prospects to access power. SF's electoral history and associated political strategy show that methods of terror have not been widely supported among the Catholic population, which circumscribes violence to extreme-ultra groups. It might be argued that the statement that the IRA/SF was making its decisions regarding prospects of gaining access to power is an oversimplification. However, if the IRA/SF were not essentially seeking power, several questions can hardly find an answer. Why did it start to participate in elections in 1982 (the expected votes were high due to the hunger strikes) when previously, running elections was considered

legitimating the "British occupation"? Why did the IRA/SF suddenly drop the policy of abstentionism in 1986, when the Republic of Ireland and Britain were reaching agreements? Why did the IRA/SF declare a ceasefire (after the unfavorable election of 1994), against the essence of its doctrine of "armed struggle"? Finally, why did it begin a decommissioning of weapons process after a successful election and when Northern Ireland was about to go back to direct rule by Britain? Why did it disarm only after it was consolidated as a major political player?

Concluding Remarks

Drawing on Hardin's rational choice approach to ethnic conflict (1995), I have attempted to trace the determinants of group conflict and mobilization in the Northern Ireland case. The focus has been on three levels: ethnic group identification, membership in the IRA, and the political leadership (SF). I have deliberately only looked at the Catholic side of the conflict for the latter two levels, that is, in the analysis of the IRA and SF.

Group identification appears to be rooted in a number of cleavages that tend to be self-enforcing, including but not limited to religion and asymmetries in economic and political power across groups. Catholics, in this sense, have incentives to be drawn together in light of the epistemological comforts of home provided by shared worldviews. The 1968 civil rights movement constituted an opportunity for the IRA to regain preponderance as a major player in the political game, capitalizing politically from "the troubles," and being able to offer its organizational skills and know-how to deal with violence. Facing a demand for defense, the IRA found room to perform its activities and push its hard-line reunification agenda. The intervention of the IRA escalated the conflict, mainly by transforming defense into systematic attack and thus fueling the retaliatory component of paramilitary violence.

Extremist's motives to engage in violence are less obvious. We have addressed evidence in the literature showing that IRA members made rational choices when joining such paramilitary force (recall Heskin and White's studies). In fact, contemporary terrorists are capable of the most precise planning and sophisticated calculations, which further undermines the extra-rational behavior account. Undoubtedly, the indoctrination process is likely key in both shaping the intensity of preferences of extremists and limiting their ability to understand the world and their own options. As Hardin points out,

> It is widely believed that such narrow views cannot be sustained by many people if they are constantly exposed to very different views. Terrorist training generally takes place in very isolated camps where there is no contrary view and where every individual is constantly reinforced in the group's belief system. (2002)

This epistemological ignorance should not be viewed as a limitation on the rationality of the extremist, given that it only represents a limitation in perceived options at the time of making a rational decision and select a course of action. If group identification was the solution to a simple coordination game, engaging in an extremist group or mobilizing in defense of group identification, is not about mere coordination anymore, since the costs of participation are high.

With regard to the political leadership, I have tried to make the argument that SF and IRA, here treated as a close unit, behaved strategically in their use of violence and threat of use of violence to acquire power. The patterns of SF's electoral results and the IRA's actions toward decommissioning reveal opportunistic behavior to maximize the chance of accessing power via support at the ballot box. As Sanders puts it "The irony of modern Northern Ireland is that a vote for Sinn Fein is effectively a vote for peace, or [...] a vote against the IRA" (Sanders 2011: 255).

Notes

1. The pluralist theory provides a convincing explanation, asserting that the cleavages are reinforcing (Rose 1971).
2. The Sutton Index of Deaths includes data since 1969 to 2001. Update of the book "An Index of Death from the Conflict in Ireland 1969–1993", CAIN Web Service of the University of Ulster at Magee. Web site: http://cain.ulst.ac.uk (Accessed October 25, 2015).
3. There is evidence indicating that Northern Ireland has become more segregated since 1998, provided that "the number of peace lines maintained by the Northern Irish Office has grown from 37 in October 2006 to 48 in November 2010" (Power 2011: 5).
4. In the Republic of Ireland, its share of votes was estimated on average to be as little as less than 2%.
5. Let me comment briefly on the issue of IRA membership. Due to the secrecy of the paramilitary activity, it is hard to learn the exact membership. The Official Northern Ireland police files indicate that 17,000 individuals have been arrested for involvement in this kind of activity (republicans and

loyalists) since 1979 (Hayes and McAllister 2001). This information is not helpful for two reasons. On the one hand, many participants might have avoided police's scrutiny. On the other hand, the police might have probably identified wrongly in many cases.

6. Catholics, as a group, cannot accurately be defined by the actions of the IRA. As Hardin (2000: 185) notes: "It is a fallacy of composition to suppose without argument that a group has the characteristics of an individual member of the group."
7. 1991 Census
8. 2011 Census for Northern Ireland, www.nisra.gov.uk (accessed on October 28, 2015)
9. Twenty-six counties formed the Irish Free State, while Northern Ireland kept the remaining six, with a Protestant majority.
10. Data calculated in base of the 1926 and 1911 census. Source: Irish Central Statistics Office, web site: www.cso.ei
11. As Hardin (2004: 180) points out: "Nationalism is a political issue only if it is intentionalist for at least many of the relevant group."
12. Even though it is true that Northern Ireland used to be economically more successful than the Irish Republic, in the past decades the performance of the latter has been better (Birnie 1998).
13. SPA was abolished in 1972.
14. We have necessarily left out other sources of hostilities between the two groups. For a detailed presentation and assessment of different accounts, see McGarry, J. and O'Leary, B. Explaining Northern Ireland: Broken Images, Blackwell Publishers Inc., Oxford, 1995.
15. See Hayes and McAllister (2001).
16. Responsible of killings: INLA (11), Real IRA (13), OIRA (8). Source: The Sutton Index of Deaths includes data since 1969 to 1998. Update of the book "An Index of Death from the Conflict in Ireland 1969–1993", CAIN Web Service of the University of Ulster at Magee. Web site: http://cain.ulst.ac.uk
17. In particular, around the time of the peak of IRA violence, there was an atmosphere at the global level that successful insurgency was a possibility.
18. Italics added.
19. According to Taylor, "volunteers had no problem in finding out how to go about joining. They would approach a senior republican in the area and drop a word on his ear" (Taylor 1999b: 89).
20. Capitals in the original.
21. The Real IRA was responsible for the Omagh bomb in 1998, an event that by then undermined the IRA's compromise to ceasefire.
22. Elections results are mainly taken from the site: http://www.ark.ac.uk/elections, a database sponsored by the University of Ulster at Magee. The

elections considered are local government, Westminster, regional, and European. The CAIN project web site has also been a source for a detailed chronology of the events that took place.

23. The Irish Free State became a Republic in 1949 and by then Dublin was presumably more concerned about their new status than about the situation in Belfast. In the middle of the tension of violent demonstrations, local IRA leadership expected to receive support from Dublin. While some arms were handled, the Irish Army did not cross the border toward Northern Ireland. In Dublin, two factors seem to have caused this refusal of support. On the one hand, the then President of Ireland, Lynch, believed that keeping distance with the IRA in Northern Ireland would lead to a voluntary withdrawal of Britain from the island. On the other hand, IRA Belfast Brigade leadership was seduced by Marxist theories, highly disapproved by Dublin (O Heithir 1997). In the end, this caused the IRA to split in late 1969 into Officials and Provisionals (Provos), opposed toward the policy in regard to the use of force and political abstention.
24. For simplicity, from now on I will use the term IRA in order to refer to the Provisional IRA.
25. Sutton Index of Deaths includes data since 1969 to 1998. Update of the book "An Index of Death from the Conflict in Ireland 1969–1993", CAIN Web Service of the University of Ulster at Magee. Web site: http://cain.ulst.ac.uk
26. Westminster proposed a different solution in 1973, which excluded paramilitary groups on both sides in the negotiations. Only the North and South government and Northern Ireland constitutional powers signed the Sunningdale Agreement, which attempted to establish "a legitimate set of governmental institutions based on 'power-sharing' and the 'Irish dimension'"(Bew et al. 1997: 39).
27. Meanwhile, bombing campaigns took place in Northern Ireland and Britain.
28. In 1985, when the British Government and the Irish Republic signed the Anglo-Irish Agreement, which enabled the latter to legitimately participate in internal affairs in Northern Ireland. This negotiation was bilateral, leaving aside once again the two paramilitary groups, and raising serious protests among Unionists.
29. They maintained the policy of abstentionism for Westminster seats.
30. The stalemate was due to the absence of loyalist paramilitary organizations in the round of negotiations.
31. "The Path to Peace", by The Irish Times, at: www.ireland.com
32. SF was now included after the declaration of cessation of fire. The same applies to loyalist extremists.
33. This arrangement has prompted Horowitz to question whether this results in the institutionalization of the opposition in the cabinet.

34. They obtained 18% of the votes.
35. There were also local government elections, but we'll focus in the Westminster election, since the local government follows the pattern marked by the Westminster elections.
36. BBC News, Northern Ireland: 1998–2001 http://news.bbc.co.uk/hi/english/uk/northern_ireland/newsid_539000/539391.stm
37. Electoral results are taken from Conflict Archive on the Internet, CAIN Project, Ulster University at Magee, Whyte, Nichollas, "Election results in Northern Ireland since 1973." External link: http://www.ark.ac.uk/elections
38. Ruohomaki, Jyrki. "Two elections, two contests: the June 2001 elections in Northern Ireland," Democratic Dialogue, August 2001. At: http://cain.ulst.ac.uk/dd/papers/elect.htm
39. BBC News, Northern Ireland: 1998–2001. IRA's statement is dated as of May 31, 2001. The elections took place on June 7, 2001. http://news.bbc.co.uk/hi/english/uk/northern_ireland/newsid_539000/539391.stm
40. Tony Blair and Bertie Ahern, respectively
41. The Guardian, August 21, 2001
42. The loyalist paramilitary groups had just been "specified" (considered their ceasefires to be at an end) by the British government.
43. There were bombings attributed to the dissident Real IRA, but the bombs were mostly planted in London, thereby not directly affecting the local Northern Ireland population.
44. Clarke, Liam and Johnston, Kathry. Martin McGuinness. From Guns to Government, Mainstream Publishing Company, 2001, Ch. 19

References

BBC News. *Northern Ireland: 1998–2001*. http://news.bbc.co.uk/hi/english/uk/northern_ireland/newsid_539000/539391.stm

Bell, J. Bowyer. 1997. *Secret Army, the IRA*. New Brunswick: Transaction Publishers.

Bew, Paul, Henry Patterson, and Paul Teague. 1997. *Northern Ireland: Between War and Peace*. London: Lawrence & Wishart Limited.

Birnie, J. Esmond. 1998. The Economics of Unionism and Nationalism. In *Rethinking Northern Ireland: Culture, Ideology and Colonialism*, ed. David Miller. New York: Addison Wesley Longman Inc.

Clarke, Liam, and Johnston Kathry. 2001. *Martin McGuinness. From Guns to Government, Ch. 19*. Edinburgh: Mainstream Publishing Company.

Conflict Archive on the Internet. CAIN Project, Ulster University at Magee. http://cain.ulst.ac.uk

Coogan, Tim Pat. 1993. *The IRA: A History*. Niwot: Roberts Rinehart Publishers.

Dimitrijevic, Vojin. 2001. "L'Eternel Retour: Terrorism-More of the Same", Feature: From Postcommunism to Post-September 11. *East European Constitutional Review* 10 (4): 84–86.

Gudgin, Graham. 1999. Discrimination in Housing and Employment Under the Stormont Administration. In *The Northern Ireland Question: Nationalism, Unionism and Partition*, ed. Patrick J. Roche and Brian Barton. Aldershot: Ashgate Publishing Ltd.

Hardin, Russell. 1995. *One for All: The Logic of Group Conflict*. Princeton: Princeton University Press.

———. 2000. Fallacies of Nationalism. In *NOMOS 42: Designing Democratic Institutions*, ed. Ian Shapiro and Stephen Macedo, 184–208. New York: New York University Press.

———. 2002. NYU Forums. Discussions and Counseling Sessions Held in Response to the Tragedy. "Terrorist Community". September 18, 2002.

———. 2004. Subnational Groups and Globalization. In *Justice and Democracy: Essays for Brian Barry*, ed. Keith Dowding, Robert E. Goodin, and Carole Pateman. Cambridge, UK: Cambridge University Press.

Hayes, Bernadette, and McAllister Ian. 2001. Sowing Dragon's Teeth: Public Support for Political Violence and Paramilitarism in Northern Ireland. *Political Studies* 49: 901–922.

Heskin, Ken. 1980. *Northern Ireland: A Psychological Analysis*. New York: Columbia University Press.

http://larkspirit.com/general/orangehist.html

Irvin, Cynthia L. 1999. *Militant Nationalism: Between Movement and Party in Ireland and the Basque Country*. Minneapolis: University of Minessota Press.

McEvoy, Joanne. 2015. *Power-Sharing Executives*. Philadelphia: University of Pennsylvania Press.

McGarry, John, and O'Leary, Brendan. 1995. *Explaining Northern Ireland: Broken Images*. Oxford: Blackwell Publishers Inc.

O Heithir, Breandan. 1997. *Pocket History of Ireland*. Dublin: The O'Brien Press.

Power, Maria. 2011. Introduction: Peacebuilding in Northern Ireland. In *Building Peace in Northern Ireland*, ed. Maria Power. Liverpool: Liverpool University Press.

Rose, Richard. 1971. *Governing Without Consensus: An Irish Perspective*. Boston: Beacon Press.

Ruane, Joseph, and Jennifer Todd. 1998. Irish Nationalism and the Conflict in Northern Ireland. In *Rethinking Northern Ireland: Culture, Ideology and Colonialism*, ed. David Miller. New York: Addison Wesley Longman Inc.

Ruohomaki, Jyrki. 2001. Two Elections, Two Contests: The June 2001 Elections in Northern Ireland. *Democratic Dialogue*, August. http://cain.ulst.ac.uk/dd/papers/elect.htm

Sanders, Andrew. 2011. *Inside the IRA: Dissident Republicans and the War for Legitimacy*. Edinburgh: Edinburgh University Press.
Surge in Support for Irish Unity, Special Report: Northern Ireland, *The Guardian*, August 21, 2001.
Taylor, Peter. 1999a. *Loyalists: War and Peace in Northern Ireland*. New York: TV Books, L.L.C.
———. 1999b. *Behind the Mask*. New York: TV Books, L.L.C.
White, Robert W. 1989. From Peaceful Protest to Guerrilla War: Micromobilization of the Provisional Irish Republican Army. *American Journal of Sociology* 94 (6): 1277–1302.
Whyte, Nichollas. *(Alliance Party of Northern Ireland) Election Results in Northern Ireland Since 1973*. http://www.ark.ac.uk/elections
Whyte, J. 1990. *Interpreting Northern Ireland*. Oxford: Clarendon Press.

Hardin's *One for All*: Insights for Human Rights

Kimberly Stanton

Introduction

Russell Hardin was a moral theorist who cared about the real world, not merely as an object of study—although he continuously drew upon historical as well as contemporary events for the examples he used to make his arguments—but because he was truly and deeply concerned about the human condition, and the prospects for improving it. His lifelong attention to the relationship between self-interest and collective action was both academic and practical: he switched from the study of mathematics and physics to political science because he felt he should be doing something useful about the Vietnam War. As the Cold War was drawing to an end, he combined writing and teaching with real-world engagement to mitigate the risk of nuclear confrontation.

In 1995 Hardin published *One for All: The Logic of Group Conflict*. As he was writing the book, the Bosnian civil war was raging; it formally ended in December of that year, a few months after the Srebrenica massacre in which more than 7000 Bosnian Muslims, mostly men and boys, were slaughtered in the space of 10 days (United Nations International Criminal Tribunal for

K. Stanton (✉)
Tom Lantos Human Rights Commission, House Committee on Foreign Affairs, Washington, DC, USA

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_12

the former Yugoslavia 2017). During the three-and-a-half-year war, more than 100,000 people were killed, of whom 35 percent were civilian men, women and children (Zwierzchowski and Tabeau 2010). The conflict was one of a series that tore apart Yugoslavia in the space of a few years, even though it had existed peacefully as a multiethnic state for decades, and its people had lived as neighbors and intermarried extensively across ethnic and religious lines.

The Yugoslavia case was one of several violent conflicts that Hardin examined briefly in *One for All*, along with Northern Ireland, Somalia, and the Hutus versus the Tutsis. He was seeking to understand "the sway of groups in our time" because their success seemed to belie the expectations of the logic of collective action, according to which self-interest generally runs counter to group interest, and as a result, people commonly fail to act collectively. How to explain, then, not only the existence of the powerful groups engaged in violent conflicts, but their persistence in spite of the often devastating consequences of the conflicts on their own members? Perhaps it was not surprising that at that time, influential observers of the terrible conflicts unfolding around the world tended to attribute the motivations for collective violence to something other than rational self-interest, such as the unleashing of primordial identities—because how else could one explain that her next-door neighbor was suddenly transformed into an enemy capable of killing her?

In the twenty-plus years since *One for All* was published, the concerns that motivated Hardin have persisted. The civil war in Bosnia was followed just a couple of years later by an armed conflict in Kosovo in which Serbian forces responded to an ethnic Albanian uprising with "a systematic campaign of terror, including murders, rapes, arsons and severe maltreatments," characterized today as ethnic cleansing (BBC News 2001). More than 13,500 people were killed, of whom more than 10,000 were civilians, and more than 1.5 million people were expelled or internally displaced, including 90 percent of the Kosovar Albanians.

More recent cases in which groups engaged in conflict have been mobilized on the basis of religious and ethnic differences include the current conflict in South Sudan, which began as a political rivalry between the new country's president and vice-president but has been transformed into a murderous civil war that breaks down largely along Nuer-Dinka lines. As many as 300,000 have been killed, 3.5 million have been uprooted from their homes, and the country is at risk of famine. The ongoing, devastating conflicts in Iraq, Syria and Yemen are all characterized by appeals to religion

and ethnicity, even though other factors, notably political power and territorial control, are clearly at play. In these wars, governments have bombed their own citizens, besieged entire communities and denied them food and medical care, while irregular armed groups have engaged in mass murder, sexual slavery and beheadings, and destroyed representations of cultural heritage. Millions have been forced to flee: the resulting humanitarian crises have reached levels not seen since the Second World War (United Nations Office for the Coordination of Humanitarian Affairs 2016).

Although I had the privilege of studying under Hardin, I did not follow his path into academics. Instead, I have worked as a practitioner in the foundation and non-profit sectors and in the United States Congress on foreign policy and human rights issues. It is with this background that recently I re-read *One for All* (Hardin 1995),[1] and discovered anew how Hardin's analysis helps explain the world's myriad conflicts, including the most violent of these. But in addition to its explanatory import, Hardin's approach has prescriptive implications. For human rights practitioners, Hardin's work can provide an analytical underpinning for human rights strategies, as well as guidance for interventions that seek to prevent or transcend violent conflict—both very welcome in a world where human rights advocates struggle to make the phrase "never again" a reality rather than a slogan.

Hardin's Approach: Key Points

In his analysis of the logic of group conflict in *One for All*, Hardin seeks to explain how an individual acting on a rational basis could come to identify with a group in a way that could motivate his or her actions; how that subjective identification with a group could be self-reinforcing over time; and, when multiple groups exist, how the norms and institutions that structure incentives could generate dynamics in which members of a group, behaving in their own self-interest, could be drawn into violent conflict, with appalling consequences. While the framework Hardin develops can be used to analyze any group interaction, the outcomes of the interactions vary: violent conflict is contingent rather than inevitable. Each step of Hardin's argument draws attention to certain empirical factors, such as the knowledge available to people, or how they are affected by the exercise of power through existing norms and institutions. For any particular case, the explanatory question becomes how the interplay of these

empirical factors with the underlying logic of group interaction affects the choices that people face.

First, Hardin makes a straightforward claim that much of our knowledge is acquired from and within a social context, and is taken on faith. Some of that received knowledge may be objectively wrong, but if that is the knowledge we have at our disposal, it will be rational to act accordingly. What we know may evolve through experience, interaction with others or our own purposeful learning, so that the knowledge on which we base rational decisions may change over time. But the central thesis is one of "common sense epistemology": what is rational in the sense of being in one's interest depends on the knowledge one has at the moment.

Second, Hardin distinguishes "identity" in an objective sense from "identification" in a subjective sense that entails commitment and motivation. A person has many objective attributes, including sex, race or ethnicity, and so has many potential sources of identification and motivation. One's commitments and motivations do not flow automatically from the objective elements of one's identity.

Third, Hardin posits that social order often evolves from interactions that take the form of coordination games, in which each party gains only if all others do as well. There is an element of accident in the choice of a specific coordination point: while everyone involved would benefit from coordinating as opposed to not, there may be no particular reason for choosing one specific coordination point as opposed to another. As a result, group coordination may turn on highly subjective considerations; elements of identity such as language, race, or ethnicity may emerge as markers that facilitate group coordination. The process of coordinating based on particular markers may be aided if a leader is urging those involved to recognize their "identity" and coordinate around it. Successful coordination may also be a tipping phenomenon—as more people act, it may make more and more sense for rational, self-interested individuals to join them. Thus Hardin argues, against the position that group identification is primordial or extra-rational, that it can be the cumulative, unintended result of many rational choices made over time.

Fourth, coordination points, once established, generate conventions that reinforce the coordinated actions over time. These conventions take the form of norms or institutions, and become sources of power by shaping and reinforcing expectations about the behavior of others, and/or because they become imbued with moral force,[2] and/or due to the creation of laws and

attendant sanctions that reinforce the benefits to an individual of behaving in keeping with the established order.

Of course, in any given time or place, more than one group is likely to exist. If the initial problem of social order was coordination based on common interest in a collective good, the later problem becomes the risk of conflict between or among groups. Hardin differentiates positional goods, like public office, from distributional goods, such as income and welfare benefits, and points out that if control of a positional good has implications for distributional goods, conflict may be exacerbated. The mechanism is discrimination: if one group gains a dominant political or economic position, it may discriminate against the other, not out of moral failing, but because once group identification exists, it will commonly be rational for members of the group to further the interests determined by that identification. When those interests have to do with resources a group has come to control, the consequences for inter-group conflict can be particularly significant.

Finally, Hardin's analysis includes a discussion of universalistic norms, those that "apply indifferently to everyone." If the motivating power of norms of exclusion derives from the incentive effects of the benefits of excluding others, but universalistic norms apply to everyone, can universalistic norms motivate individuals to act? His answer is yes, in part, but not very forcefully. He argues that universalistic norms, except for those governing essentially dyadic, ongoing relationships such as promise-keeping, truth-telling, and fidelity among close associates, are generally weak, and are not well reinforced by incentives of self-interest.

Hardin develops each of these points—the nature of knowledge, the process of group identification, the coordination process that generates norms and institutions, the risk of discriminatory use of power, and the weakness of universal norms—as part of his explanation of how rational individuals can fall into violent conflict. But the argument also suggests possible points of intervention to prevent or transcend violent conflict. If knowledge is limited, can it be expanded? Can subjective group identification be modified? Can discriminatory norms and institutions be modified? And if so, how? Both the explanatory and the prescriptive implications are relevant to the work of human rights practitioners, who proceed from the premise that the protection, promotion and defense of human rights are both ends in and of themselves, and key for preventing violent conflict and for redressing its consequences.

The Strategic Logic of Human Rights Activism

Human rights norms are purposefully universal. A clear foundational statement is found in the Universal Declaration of Human Rights (UDHR), adopted by the United Nations General Assembly in 1948, which begins by recognizing the inherent dignity and the equal and inalienable rights of "all members of the human family." Every article either states a principle that applies to "everyone," or identifies forms of treatment that "no one" shall be subject to (Universal Declaration of Human Rights 1948). The stated rights are broad and encompassing: life, liberty, security of person, equality before the law, due process, privacy, freedom of thought and belief, nationality, freedom of movement, equal pay, just remuneration, adequate standard of living, participation in cultural life, and more.

But the most cursory review of the news most mornings is sufficient to conclude that the universalist aspirations embodied in human rights standards do not match people's lived reality. Government entities, UN bodies, and independent non-governmental human rights organizations regularly document serious to severe restrictions on the exercise of a broad range of rights.[3] Freedom House, which compares the state of fundamental civil and political rights over time, has identified an overall deterioration in respect for fundamental human rights in recent years, even in formally democratic regimes (Freedom House 2017). The violence unleashed by groups motivated by norms of exclusion, and the failure of state responses, are such that Zeid Ra'ad Al Hussein, the U.N. High Commissioner for Human Rights, is actively warning that commitments to human rights norms are being dismantled (MSN 2017).

Trying to close the gap between aspiration and reality is the stuff of human rights activism. Often human rights work is perceived as idealist or morally-driven,[4] and sometimes undervalued or even dismissed by those who see the world in something akin to Hobbesian terms. But in fact, human rights work has an underlying strategic logic that is precisely what Hardin's analysis in *One for All* would lead us to expect. The fundamental goal of human rights work is to create conditions that permit individuals to live their lives fully and without violence—to change existing coordination points, if you will, in keeping with its universalist norms. Human rights work is meant to restructure established power relations, which is why it so often encounters fierce resistance.

How do human rights practitioners set about upending power relations? One important part of their work is to educate people about their human

rights. Examples range from short introductions to human rights designed for local communities or grassroots advocates, to online practitioner courses, to university courses incorporated into larger programs of study, to full graduate degrees. In some locales, human rights are incorporated into public school curricula even at the elementary level. Hardin's analysis in *One for All* suggests at least two ways that learning about human rights might affect the motivational power of universal human rights norms.

One would be by affecting the everyday knowledge the people rely on for making rational decisions—by shaping the content of commonsense epistemology. To the extent that universal human rights norms—such as admonitions against discrimination on the basis of sex, or race, or disability, and the very idea of equality—are inculcated at the household or community level, it becomes more likely that they will shape and motivate individual behavior. These kinds of principles can be reinforced in dyadic interactions at the household, school or community level, and will become more deeply embedded to the extent that they are reinforced by other sources of knowledge, such as learning based on observing the behavior of others.

Second, human rights education can contribute to the evolution of individuals' knowledge. As Hardin observers, while each of us begins with received knowledge, what we know and thus the basis on which we make rational decisions evolves over time, in part due to purposeful learning. In this sense, human rights learning is like any other in its potential to change the way we understand our options and frame our choices. Learning about human rights may not guarantee that someone will always act in ways that are consistent with respecting human rights. But it can provide new information, encourage reflection and shape knowledge in ways that may alter a decision-making calculus, for example, by suggesting a broader range of options than might otherwise have been perceived.

At first glance, human rights education might not seem to be a particularly powerful strategy for changing the world. But it can work on at least a couple of levels: it can be very empowering for individuals in their daily lives, and when a sufficient number of people are reached, expectations of just treatment can tip in a new direction, with consequences for social and political action—as happened with the civil rights movement in the United States in the 1960s. The fact that human rights defenders are so often imprisoned or killed suggests that many governments do fear the power of a universalist rights message[5]—even as that same mistreatment may

reinforce perceptions of injustice and further strengthen the appeal of human rights.

Another way in which human rights practitioners seek to shift the calculus of the powerful in favor of universalist human rights norms is by increasing the costs of non-compliance. It is because human rights practitioners are well aware that human rights norms are difficult to enforce that so much of their attention has been dedicated to transforming the aspirational norms into law and developing enforcement mechanisms.

Because the UDHR was not legally binding, its approval was followed by years of work to produce the two core human rights conventions, the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social and Cultural Rights (ICESCR), both universal in scope.[6] Subsequently, the core human rights norms have been further developed and extended through additional, complementary conventions that, although more specific in their focus, build upon and reaffirm the universal principles asserted in the core rights instruments.[7] As treaties, once acceded to or ratified, the conventions are legally binding under international law. This is not the case for declarations or resolutions, which often precede new treaties.[8] Once a consensus is built around the need for new human rights protections, the goal tends to be to achieve a new treaty, precisely because of its legally binding character.

Once a human rights convention comes into existence, the next goal becomes to ensure the broadest possible adherence—ideally, universal. A minimum number of states party is necessary for any treaty to enter into force.[9] Beyond that, the logic behind campaigns for ratification is clear: the larger the number of countries that adhere, the stronger the shared expectations of compliance and, in principle, the more consolidated the norms. The United Nations includes 197 countries and tracks 18 human rights treaties (of which 9 are optional protocols) in its ratification database.[10] The convention that comes closest to universal ratification is the Convention on the Rights of the Child, with 196 states party, followed by Convention on the Elimination of All Forms of Discrimination Against Women, with 189. The weakest convention by this measure is the International Convention on the Rights of All Migrants Workers and Members of Their Families, which has only 51 states party. It would be fair to conclude that the standards codified in this Convention are far from universally accepted at the international level.

Hardin emphasizes the role that sanctions play in reinforcing compliance with norms. While adherence to a human rights convention by a large

number of countries should strengthen the related norms and contribute to fostering compliance, it does not substitute for an enforcement mechanism that can increase the costs of non-compliance. Finding ways to sanction non-compliance is another priority for human rights practitioners.

One clear strategy has been to advocate for formal enforcement mechanisms directly linked to the conventions. Some of these do exist, although they often take the form of complaint mechanisms rather than judicial processes,[11] and they are optional—countries must choose to submit to their jurisdiction. Even when rights conventions themselves have achieved near universal adherence, that has not carried over to the related enforcement mechanisms: for example, the ICCPR has 169 states party, but only 116 have ratified or acceded to the Optional Protocol, which allows the Human Rights Committee created by the Covenant to consider communications from individuals claiming to be victims of violations. For the ICESCR, the situation is far worse: the convention has 165 states party, but the Optional Protocol has only attracted 22. The lower adherence to enforcement mechanisms likely has several explanations, ranging from hypocrisy to a genuine concern for their political misuse. But even though these mechanisms can be fairly characterized as relatively weak, human rights advocates will continue to push for them because some purchase on accountability is always better than none.

At the same time, an increasingly important and promising strategy for reinforcing human rights norms through enforcement is through the incorporation of the norms into domestic law. This is the ongoing process of bringing domestic law into conformity with international conventions once they have been ratified. It may occur through litigation, legislation, or legal interpretation, and in turn contribute to making operational the principle of complementarity, according to which courts at the national level should have the first opportunity to address cases of human rights violations. As national jurisdictions integrate international human rights standards into their laws and jurisprudence, enforcement of the international norms is transferred to the domestic judicial system, which in principle will be better positioned to attribute responsibility for violations and will likely adjudicate more rapidly than international complaint mechanisms.

Finally, when formal mechanisms do not exist, rights practitioners seeking to enforce a convention develop other strategies to try to ensure compliance, such as shaming, or the creation of a sanctions mechanism that is independent of the convention itself, such as a regime of economic sanctions.[12] Shaming works best when states have become party to a rights

treaty, or include human rights language in their domestic law and constitutions, because both create a basis for demanding accountability. Economic sanctions are imposed externally by other countries that have signaled that they share human rights norms, and so can be pressed into taking action. Both of these strategies generally require that human rights practitioners mobilize political support at the popular and elite levels, which is more easily done when the relevant actors believe in human rights and are sensitive to their own reputation for compliance. Enforcement thus circles back to knowledge and subjective motivation.

Let me return to the problem that motivated Hardin to write *One for All*, violent group conflict. The violations that occur as a result of violent conflict are well-known and often highly visible: the indiscriminate killing of civilians, extra-judicial executions, sexual violence, forced recruitment of minors, forced displacement. But the role of human rights violations in generating violent conflict can be less evident, if only because widespread patterns of human rights violations have been known to exist for long periods of time without any apparent political consequence.

Hardin's analysis of the dynamics that generate violent conflict does not depend on the existence of widespread, multiple human rights violations. Only one human rights violation appears to be necessary: discrimination based on group membership. It is one group's (rational) pursuit of its interests at the expense of another's that can generate resentments and provide a basis for mobilization, especially if the group gains a dominant political or economic position that entails the control of resources. All of the cases of violent conflict that Hardin analyzes in *One for All* involve discrimination in this sense by a dominant group against others.

Discrimination is often a built into a particular social order. While Hardin is persuasive in arguing that much of the social order likely evolved from interactions that took the form of coordination games, it is clear that this is not always the case. Sometimes order in the form of an institutional arrangement is imposed, and then sustained over time, in the absence of acquiescence by a significant minority, or even a majority, through a combination of measures that include the use of force. In such cases, discrimination may be a feature, and may contribute to transforming elements of objective identity into subjective group identification and generating a sociopolitical dynamic that risks conflict.

Northern Ireland, a case that Hardin included in *One for All*, provides a clear example. When the British partitioned Ireland in 1920, the borders were drawn in such a way as to ensure a Protestant majority in the north.

When the southern partition became independent a year later, the north, with a large Catholic minority, opted out. Education, neighborhoods, workplaces, entertainment and many other social activities were segregated between the Protestants and the Catholics. The names of places denoted religious and national affiliation. There was political underrepresentation of Catholics in the Northern Ireland parliament, the Stormont, and Catholic voting rights were restricted. The police force, the Royal Ulster Constabulary, was legally required to reserve one-third of its spots for Catholics, but the number of Catholics never exceeded 12 percent. These instances of social and political discrimination co-existed with economic discrimination, and persisted for decades, more than enough time to generate subjective group identification. When Catholics who protested for civil rights and an end to discrimination were repressed in 1968 and avenues for non-violent change closed off, it became possible, perhaps even relatively easy, to mobilize people around a nationalist identity.

In the twenty-first century, there are innumerable examples not only of the existence of laws, policies, and institutions that are deeply discriminatory against specific populations, but of their purposeful employ in furtherance of the agendas of dominant groups in circumscribed territories. Dominance may correspond with majority status, as it did in the United States during the civil rights era, and as is the case today for the Chinese vis-à-vis the Tibetans, or the Indian government vis-à-vis any number of religious minorities. But at least as often, the dominant group is a minority, or risks becoming one, as is the situation for the Sunni government of Bahrain vis-à-vis its majority Shi'a population, or the Israelis against the Palestinians if the option of a two-state solution disappears. In these kinds of situations, the egalitarian, universalist commitments of human rights practitioners are especially threatening because their fulfillment, almost by definition, challenges the existing configuration of power.

In human rights law, the guarantees of equality and non-discrimination are fundamental, legally binding obligations. Combatting racial and religious discrimination, and discrimination against women, indigenous peoples, migrants and other minorities, is an overarching priority of the United Nations human rights system, which rightly links situations of discrimination to grave human rights abuses (Cf. United Nations Office of High Commissioner of Human Rights 2017). The actions that human rights practitioners undertake to combat discrimination overlap substantially with recommendations that can be derived from Hardin's analysis of the dynamics that lead to violent conflict. For instance, laws, policies or

institutions that have the effect of reinforcing differences among religious or ethnic groups should be reformed. Security forces should be integrated. In a diverse country, the composition of personnel in institutions with the power to distribute resources should reflect that diversity. Institutional guarantees for equity and equality should be created or strengthened. From a human rights perspective, recommendations like these are designed to further guarantees of equality under the law. From a Hardin perspective, these kinds of steps are meant to reduce the impact of objective markers of identity on people's life prospects, and similarly reduce the benefits from subjective group identification, weakening its power to motivate.

In the end, Hardin's analysis in *One for All* is not reassuring with regard to the enormity of the challenge presented by violent group conflict. But his explanation of its dynamics does suggest a sort of a roadmap for proceeding by pointing us to objective factors that, to the extent they are subject to intervention, could make a positive difference—because in the end, human beings are rational actors, so changing the incentives they face can make a world of difference. This is what the human rights movement is also about, whether or not we always fully realize it.

Notes

1. I re-read *One for All* in preparation for the Festschrift held in Hardin's honor at New York University in November 2015. This article has its origins in that event.
2. But Hardin argues strongly against the is-ought fallacy. The existence of a particular coordination point cannot be taken per se to imply its moral rightness. Cf. Hardin (1995, Chap. 3, p. 60ff).
3. See for example the annual U.S. Department of State Country Reports on Human Rights Practices (2017).
4. Hardin does not deny the possibility that moral or philosophic beliefs can motivate individual action. He simply argues that such action is unlikely to be widespread unless it coincides with self-interest.
5. For current information on threatened human rights defenders around the world, see https://www.frontlinedefenders.org, accessed May 21, 2017.
6. The texts of the ICCPR and the ICESCR are available on the website of the UN Office of the High Commissioner for Human Rights (OHCHR): http://www.ohchr.org/EN/ProfessionalInterest/Pages/UniversalHumanRightsInstruments.aspx, accessed May 21, 2017.
7. Some of these address the situation of specific populations, such as the Convention on the Elimination of All Forms of Discrimination Against

Women (CEDAW), while others refine and elaborate certain rights, such as the Convention Against Torture (CAT), or prohibitions, such as the Convention for the Protection of All Persons from Enforced Disappearance. The preambles of more recent conventions always make reference to the universal principles of the UDHR.

8. For example, several declarations preceded the International Convention on the Elimination of All Forms of Racial Discrimination and are cited in the Preamble to the Convention. The text of the Convention is available on the website of the Office of the UN High Commissioner for Human Rights: http://www.ohchr.org/EN/ProfessionalInterest/Pages/CERD.aspx
9. Each convention specifies the minimum number of states party that are required for it to enter into force. For the ICCPR and the ICESCR, the number was 35. For the Convention on the Rights of Persons with Disabilities, the number was 20.
10. Available at http://indicators.ohchr.org, accessed May 21, 2017.
11. Exceptions include the regional human rights courts, in particular the European Court of Human Rights and the Inter-American Court of Human Rights, as well as the International Criminal Court, established by the Rome Statute, which investigates and prosecutes individuals accused of genocide, war crimes and crimes against humanity. In all cases, countries must choose to accept the jurisdiction.
12. The Enough Project is an example of a human rights organization that advocates a form of accountability for human rights violations by promoting economic sanctions. Cf. http://www.enoughproject.org, accessed May 21, 2017.

References

BBC News. 2001. *Kosovo Assault 'Was Not Genocide'*. September 7, 2001. http://news.bbc.co.uk/2/hi/europe/1530781.stm. Accessed 17 May 2017.

Freedom House. *Freedom in the World*. Available for Various Years at https://freedomhouse.org/report-types/freedom-world. Accessed 21 May 2017.

Hardin, Russell. 1995. *One for All*. Princeton: Princeton University Press.

MSN Video News Interview, April 5, 2017. Available at https://www.msn.com/en-gb/news/video/un-human-rights-chief-commitments-to-rights-norms-are-being-dismantled/vi-BBAJGcM. Accessed 21 May 2017.

U.S. Department of State. 2017. *Country Reports on Human Rights Practices*. Available at https://www.state.gov/j/drl/rls/hrrpt/. Accessed 21 May 2017.

United Nations International Criminal Tribunal for the Former Yugoslavia. 2017. *ICTY Remembers: The Srebrenica Genocide 1995–2015*. http://www.icty.org/specials/srebrenica20/?q=srebrenica20/. Accessed 21 May 2017.

United Nations Office for the Coordination of Humanitarian Affairs. 2016. *Record International Humanitarian Appeal Requires $22.2 Billion for 2017*. December 5, 2016. https://www.unocha.org/story/record-international-humanitarian-appeal-requires-222-billion-2017. Accessed 21 May 2017.

United Nations Office of High Commissioner for Human Rights. 2017. *A Special Focus on Discrimination*. Available at http://www.ohchr.org/EN/Issues/Discrimination/Pages/discrimination.aspx. Accessed 21 May 2017.

Universal Declaration of Human Rights. 1948. Available at http://www.un.org/en/universal-declaration-human-rights/index.html. Accessed 21 May 2017.

Zwierzchowski, Jan, and Ewa Tabeau. 2010. *The 1992–1995 War in Bosnia and Herzogovina: Census-Based Multiple System Estimation of Casualties Undercount*. February 1, 2010. http://www.icty.org/x/file/About/OTP/War_Demographics/en/bih_casualty_undercount_conf_paper_100201.pdf. Accessed 21 May 2017.

Norm-Supporting Emotions: From Villages to Complex Societies

Cristina Bicchieri and Erik Thulin

How do socially imposed rules develop into internalized pro-social codes? In the article "From Bodo Ethics to Distributive Justice" (Hardin 1999), Russell Hardin discusses one of the central themes of his work: How we "export" social order from a small, insular community to a large, anonymous society. In Bodo's small village, everyone knows everyone else, interactions are face-to-face, and people live relatively isolated from other communities. In this context, the social norms developed by the community are easily enforceable. But what about large, anonymous societies, where monitoring is difficult and costly and sanctioning transgressions carries a greater risk? When unobserved, only someone with an inner motivation to behave in a socially beneficial way will continue to obey the informal rules. How such an inner motivation develops is a topic of debate in moral philosophy and psychology, especially whether pro-social decisions are a matter of rationality or are driven by emotions. Supporters of the emotional drivers of pro-social behavior argue that anger and empathy play an essential

C. Bicchieri (✉)
Philosophy and Psychology Departments, University of Pennsylvania, Philadelphia, PA, USA

E. Thulin
Psychology Department, University of Pennsylvania, Philadelphia, PA, USA

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2_13

role (Fessler and Haley 2003; Haidt 2003). In our subsequent discussion, we will focus on the role of these emotions in compensatory and sanctioning behavior.

Participants in laboratory games become angry at those who free-ride on the public good (Fehr and Gachter 2004) or offer an unfair deal (Pillutla and Murnighan 1996). Often, this anger leads to action. Experimental work on punishment has shown that—at least in the relatively aseptic lab environment—people are willing to bear a cost to punish norm violators (Fehr and Gachter 2004). Even more telling, people become angry as third-party observers of unfair treatment and engage in what is known as altruistic third-party punishment (Nelson and Zeelenberg 2009; Fehr and Fischbacher 2004). There are a variety of proposed drivers connecting anger and punishment behavior. For example, we may punish to give the perpetrator their "just deserts" (Carlsmith et al. 2002), as an act of revenge on behalf of the victim. Alternatively, punishment may be driven by a desire to restore the values of the community (Wenzel and Thielmann 2006). In this case, observers see the offense as questioning the social norms of the community, and punishment serves to reassert those violated rules. Anger, in other words, serves to restore the moral balance by bringing the perpetrator down from the ill-gotten position. Hardin would probably suggest, in contrast, that anger allows the punisher to strengthen her own reputation in the community. In the short run, engaging in costly signaling is not a rational strategy, especially when the punisher is a third party not directly involved in the interaction. Yet in the long run this behavior may help create a good reputation for the punisher. Our view is that emotional reactions play an important role in supporting social norms, but these emotions are strictly dependent on the agents' expectations. These shared expectations, in turn, are constitutive of social norms (Bicchieri 2006).

Many of the reasons for punishing norm violators also apply to compensating the victims of norm violations. Just as punishment restores the moral balance by bringing the perpetrator down from an ill-gotten position, victim compensation restores the moral balance by returning the victim to their original state. Similarly, just as punishing a perpetrator reasserts the value of the social norm through costly signaling, so does the costly signal of compensating the victim of the violation. Given this symmetry in context and motivation between the punishment of norm violators and the compensation of their victims, we propose that the emotions preceding these behaviors may be more similar than previously thought.

Much of past research on helping behavior has focused on empathic concern, a constellation of emotions including feelings of sympathy, compassion, and tenderness as the primary driver of helping behavior (Coke et al. 1978; Batson 1981; Toi and Batson 1982). Empathic concern appears to lead to helping someone who receives an unequal allocation in an economic game (Leliveld et al. 2012). Anger, on the other hand, is associated with the willingness to punish. Though the conventional view is that punishment and compensation of victims of norm violations are driven by very different emotional states, our work has shown that moral outrage, a measure of anger, drives both behaviors under specific conditions.

Emotions are closely connected with cognition. Therefore, a coarse assignment of emotion to a specific behavior (empathy/compensation and anger/punishment) does not do justice to the subtle interaction between beliefs and emotional states. Just as responders in Ultimatum games do not get angry when their offer is the outcome of a random device, compensators are apt to distinguish between different causes of harm, and their emotional reactions attest to these distinctions. We have recently shown that the compensation of the victim of a norm violation, a behavior observed in recent work (Charness et al. 2008; Chavez and Bicchieri 2013), is driven by the compensator's feeling of moral outrage, rather than their empathic concern for the victim (Thulin and Bicchieri 2015). We shall report here the results of experiments we conducted to test the hypothesis that moral outrage (anger) drives the compensation of victims, but only when the victim's loss was the result of a social norm violation. In other cases, empathy drives victim's compensation (Thulin and Bicchieri 2015, 2017).

We ran three studies to investigate the hypothesis that compensation, in the case of a norm violation, is driven by moral outrage, the same emotion that drives third-party punishment. In Study 1 (Thulin and Bicchieri 2015), we looked at the relationships between one's general dispositions to feel moral outrage and empathic concern (trait moral outrage and trait empathic concern), and the willingness to compensate. Participants were given the opportunity to compensate a player's loss in a Trust game. In this study, we confirmed our hypothesis that trait level moral outrage predicts participants' willingness to compensate the victim of a social norm violation, but not when the loss was due to an investment gone awry or random chance.

In Studies 2 and 3 we aimed to make a stronger causal claim for the role of moral outrage in driving compensation of norm violation victims. In Study 2, we experimentally manipulated participants' empathic concern and moral outrage. In Study 3, we successfully extended and replicated Study

2 using monetary incentives and a simplified compensation dependent measure. In both studies, we validated our prediction that increasing moral outrage would increase the compensation of norm violation victims, but not of those who experienced a loss for other reasons (Thulin and Bicchieri 2017).

Study 1

In this study, we assessed the relationship between moral outrage and compensation at the trait level. We measured participants' willingness to compensate across a variety of hypothetical contexts. In each context, another person lost money, either due to someone else's violation of a social norm, a bad investment, or random chance. After observing this person losing money, the participant had the opportunity to compensate this person for their loss by transferring some of their own endowment to that person. We then measured each individual's general propensity to feel both moral outrage and empathic concern. We predicted that one's propensity to experience moral outrage would contribute to their willingness to compensate beyond their propensity of experience empathic concern, but only in the context of a norm violation.

Method

We recruited 241 participants (108 men, mean age of 33) from the Amazon's Mechanical Turk (AMT) platform to participate in this study. We chose AMT to draw a more diverse sample than available from the traditional undergraduate population. Previous work found that AMT samples are more diverse on age, geography, and ethnicity than undergraduate populations. In addition, the same work found AMT to be at least as reliable as that gathered through traditional methods (Buhrmester et al. 2011).

To measure the degree to which empathic concern and moral outrage influenced third-party willingness to compensate across a variety of contexts, we used a series of modified hypothetical trust games with third-party compensators. In the original trust game, the experimenter assigned participants to one of two roles, either that of the investor or the trustee. The investor received an initial endowment, and could choose to transfer any of that amount to the trustee. The experimenter would then triple any amount transferred by the investor. The trustee could then choose to send back any portion of the tripled amount to the investor.

Using the original trust game as a foundation, we created three different interactions in which a participant may lose their endowment due to either the violation of a reciprocity norm, a bad investment, or random chance. These three situations served as three conditions in the study.

In the norm violation interaction, the experimenter endowed an investor with $10. The investor could then choose whether or not to transfer that $10 to the trustee. If they chose to keep the $10, the game ended. If they chose to transfer the $10 to the trustee, the experimenter quadrupled the amount to $40. At this point, the trustee could then choose to either keep the $40 or to return half ($20) to the investor. If the trustee chose to return half, the interaction ended. However, if the trustee chose to keep the entire $40, a third-party observer, who was endowed with $10, was given an incentive-compatible elicitation measuring the most they would be willing to pay to restore the investor to their original $10.

The bad investment situation was very similar to the norm violation situation, but with a single modification. Instead of the trustee having a *choice* of whether to transfer back $20 of the $40 transfer to the investor, a randomizing device selected whether to return the $20. We chose the probabilities of an 80% chance of return of the $20 and a 20% chance of returning $0, which was known to all participants. These values were chosen to mimic the return rates in trust games of a similar setup (Fetchenhauer and Dunning 2009). After observing the interaction, if the $20 was not returned to the investor, the third-party observer had the same choice as in the norm violation condition. Importantly, in this version of the interaction, if the investor chose to transfer their endowment, whether or not the $20 was returned to them no longer depended on the trustee conforming to a norm.

Finally, the random chance interaction was similar to the bad investment interaction, but with one more modification. Instead of the investor having the *choice* of whether their $10 is transferred to the trustee (and then quadrupled by the experimenter), a randomizing device selected whether the $10 is transferred. We chose probabilities of a 50% chance of transferring the $10 and a 50% chance of not transferring the $10, which was common knowledge. These were again chosen to mimic the investment rates in trust games in a similar setup (Fetchenhauer and Dunning 2009). As in the bad investment interaction, if the $10 was transferred to the trustee, a randomizing device then selected whether or not $20 is returned to the investor. If the $20 was not transferred back to the investor, the third-party observer then had the same choice as in the previous two situations.

We randomly assigned each participant to one of the three interactions. After reading the complete description of one of the interactions, each respondent participated in a hypothetical instance of the interaction as the third-party observer in which the investor's money was transferred to the trustee, but none was returned to the investor. After answering how much they would be willing to pay to restore the investor to their original $10, participants responded to inventories of trait propensity to feel moral outrage and trait propensity to feel empathic concern.

We adapted the four item trait moral outrage scale from previous work (Wakslak et al. 2007). For each item, participants expressed their agreement with a statement on a 7-point scale from "does not describe me well" to "describes me very well". Example statements included "I feel angry when I learn about people suffering from unfairness" and "I think it's shameful when injustice is allowed to occur".

We used the seven item Empathic Concern Subscale of the Interpersonal Reactivity Index to measure trait empathic concern (Davis 1983). For each item, participants expressed how well it described them, on a five point scale of "does not describe me well" to "describes me very well". Items included "Sometimes, I don't feel very sorry for other people when they are having problems" and "I often have tender, concerned feelings for people less fortunate than me".

Results

188 participants (78% of the sample) correctly responded to at least nine of the ten comprehension questions asked throughout the instructions. In order to ensure high quality data, we used this subset in further analyses.

We found the four item trait moral outrage scale and seven item trait empathic concern scale to be highly internally reliable ($\alpha = .91$ and $\alpha = .90$, respectively). Additionally, trait level empathic concern and moral outrage were highly correlated with each other $r(186) = .62$, $p < .001$. This high degree of correlation leads us to conduct all analyses of these variables controlling for the other in order to isolate the unique contribution of each.

For each condition, we analyzed the partial correlation between compensation and trait moral outrage controlling for trait empathic concern as well as trait empathic concern controlling for trait moral outrage. These results can be found in Table 1.

Our key prediction was that moral outrage would predict compensation in the Norm Violation condition, while not doing so in the Random Chance

Table 1 Partial correlations between compensation in each condition and empathic concern controlling for moral outrage and moral outrage controlling for empathic concern

Condition	*Empathic concern*	*Moral outrage*
Norm violation	−.152	.270*
Bad investment	.397*	−.092
Random chance	.012	−.063

Note. All values are partial Pearson correlation coefficients
*$p < .05$

and Bad Investment conditions. We see this supported in the Moral Outrage column of Table 1. Here we see that, controlling for empathic concern, moral outrage significantly correlated with compensation in the Norm Violation condition $r(65) = .27$, $p = .027$. Also importantly, we see that, controlling for empathic concern, moral outrage predicted compensation in neither the Random Chance condition $r(56) = -.063$, $p = .636$ nor the Bad Investment condition $r(58) = -.091$, $p = .49$. In fact, both of these non-significant effects had a negative sign.

We observed that empathic concern, controlling for moral outrage, was correlated with compensation in the Bad Investment condition $r(58) = .40$, $p = .002$. However, empathic concern was not correlated with compensation in either the Norm Violation condition $r(65) = -.15$, $p = .22$ or the Random Chance condition $r(56) = .01$, $p = .93$.

Discussion

Past research includes numerous examples of helping behavior correlating with empathic concern, across a variety of contexts, from volunteering to help a sick student to paying to compensate someone who received an unfair allocation in a behavioral game (Coke et al. 1978; Leliveld et al. 2012; Toi and Batson 1982). Our initial finding that the dispositions to feel moral outrage and the disposition to feel empathic concern are highly correlated suggests an important caveat when interpreting previous findings: as these studies did not address moral outrage as a covariate, it is possible that effects interpreted as being driven by empathic concern may in fact have been driven by an important third variable, namely moral outrage. Study 1 investigated the plausibility of this claim, looking at the unique contributions of trait empathic concern and trait moral outrage across three contexts.

In support of this past literature, we find that empathic concern *does* maintain a unique correlation with compensation controlling for moral outrage, but only in particular contexts, namely in the Bad Investment condition where someone makes a risky decision and suffers a loss. We do not see any unique correlation between compensation and empathic concern in the Random Chance condition, where all transfers were randomized. Although not directly linked to the questions at hand, future work may illuminate what differences between the Bad Investment and Random Chance conditions lead to the differing effect of empathic concern, and perhaps answer what motivations may be present in compensating the victims in a random chance-like scenario.

Our focal question for this study asked whether a propensity to feel moral outrage was related to a willingness to compensate, and whether that effect was limited to the case of social norm violations. The analysis of the correlations of moral outrage with compensation, controlling for empathic concern, across the various conditions suggests the answer to both questions is yes. In the case of the social norm violation, we find that compensation correlated with moral outrage, controlling for empathic concern. In addition, we find that moral outrage did not correlate with compensation in the other two conditions.

Study 2.a

Study 2 extends the findings of Study 1 from the trait domain into that of emotional states. Study 2.a investigates the relationship between participants' current level of moral outrage and the degree to which they are willing to compensate. Whereas the previous study relied on correlational relationships with trait variables, we are able to manipulate emotional states, allowing for stronger causal claims. In this study, we manipulate the amount of moral outrage a participant experiences using video inductions. We then assess their willingness to compensate across the three hypothetical situations used in Study 1. Finally, we measured the degree to which each participant was currently experiencing moral outrage and empathic concern. We predicted that those led to experience high moral outrage would be willing to compensate more than those who were not, but that this effect would be limited to the norm violation context. Additionally, we predicted that, controlling for empathic concern as a covariate, experienced moral outrage would mediate the effect of the video induction on willingness to compensate.

Method

We recruited 990 participants (471 men, mean age of 33) from the AMT platform to participate in this study. We experimentally manipulated moral outrage, measuring its effect on compensation across the three hypothetical situations developed in Study 1: Social Norm Violation, Bad Investment, or Random Chance. Each participant read instructions describing the interaction, while answering a series of comprehension questions throughout. After reading the instructions but before being told what role in the interaction they would be assigned to, participants watched a short video, serving as the manipulation of moral outrage. This manipulation takes advantage of people's tendency to attribute arousal states such as anger to whatever stimulus they are currently being exposed to (Schachter and Singer 1962). Those assigned to high moral outrage watched a short video of a boy being attacked by a bully, which past work identified as significantly increasing moral outrage while having minimal effect on other emotions (Lerner et al. 1998). Participants assigned to low moral outrage watched a video of abstract line patterns, previously found to be emotionally neutral (Gross and Levenson 1995).

After watching one of the two videos, all participants were assigned to the role of the third-party observer and asked how they would respond if the investor's funds were transferred to the trustee, but none were returned to the investor. Participants were then given the same hypothetical version of an incentive-compatible elicitation used in Study 1, measuring their willingness to pay to restore the investor to their original $10. After giving their responses, participants then answered a series of questions measuring their current levels of empathic concern and moral outrage.

The four item state moral outrage scale was used in previous work for the same purpose (Piazza et al. 2013). For each item, participants rated the degree to which they agreed with the item on a five-point scale from "Strongly Agree" to "Strongly Disagree". Example items included "I feel angry" and "I feel outraged". We adapted three items from the Empathic Concern Subscale of the Interpersonal Reactivity Index used in Study 1 in order to measure state empathic concern. For each item, participants rated how much they agreed with the item on the same five point scale used for the state moral outrage items. Items included "I feel sorry for Person A" and "I was disturbed by what happened to Person A", Person A being the investor in their interaction.

Results

754 (76%) of participants correctly responded to 9 of the 10 comprehension questions. To ensure data quality, we only analyzed the responses from these participants.

Both the four item state moral outrage scale and the three item empathic concern scale showed high degrees of internal reliability ($\alpha = .958$ and $\alpha = .847$, respectively). Using the moral outrage scale as a manipulation check, we found that moral outrage was significantly manipulated in the norm violation situation $t(230) = 5.18$, $p < .001$, random chance situation $t(262) = 4.51$, $p < .001$, and the bad investment situation $t(256) = 3.27$, $p = .001$. These effects ranged in size across situations from $d = .41$ to $d = .68$, demonstrating a medium sized effect of the video manipulation on moral outrage.

We report mean levels of compensation across conditions in Fig. 1. In the norm violation situation we found that those who watched the moral outrage video (M = 4.09) were willing to pay significantly more to compensate than those who watched the neutral control video (M = 3.16) $t(230) = 2.41$, $p = .017$. We then tested whether a subject's feeling of moral outrage mediated this effect, the results of which can be found in Fig. 2. In this and all following mediation analyses, we ran non-parametric biased corrected bootstrap analysis (Preacher and Hayes 2008) with 10,000

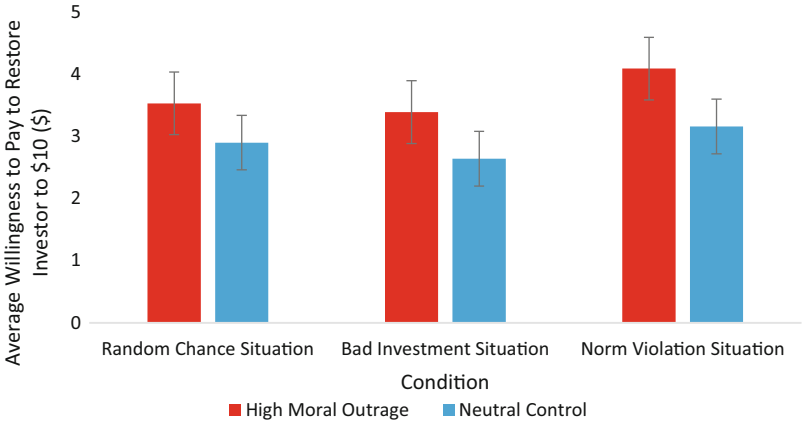


Fig. 1 Average willingness to pay of participants to restore investor to $10 by condition

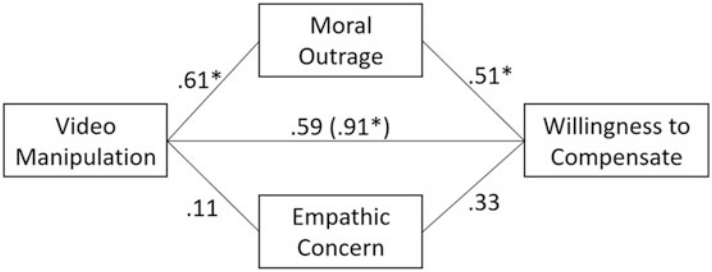


Fig. 2 Standardized regression coefficient for the relationship between the video manipulation and willingness to compensate, mediated by moral outrage and empathic concern (*$p<.05$)

resamples. Controlling for empathic concern as a covariate, moral outrage significantly mediated the effect of the video manipulation on the amount participants were willing to pay to compensate $B = .31$, 95% CI = .08 to .65.

Participants who watched the moral outrage-inducing video in the bad investment situation (M = 3.53) also compensated significantly more than those who watched the neutral control video (M = 2.64) $t(262) = .42$, $p = .042$. We observed that, in that situation, empathic concern also differed significantly between the moral outrage and control video conditions $t(262) = 2.40$, $p = .017$. Mediation analysis showed that while moral outrage was *not* a significant mediator of the effect of the video on compensation ($B = .01$, 95% CI = −.22 to .21), empathic concern *was* a significant mediator $B = .19$, 95% CI = .02 to .38.

In the random chance situation, those who watched the moral outrage video (M = 4.09) also compensated significantly more than those who did not (M = 3.16) $t(256) = 2.39$, $p = .017$. However, similar to the bad investment situation, controlling for empathic concern as a covariate, moral outrage was *not* a significant mediator of the effect of the video on compensation $B = -.17$, 95% CI = −.25, .07.

Discussion

The finding that increased moral outrage lead to increased willingness to compensate in the norm violation situation provides support for the causal role of moral outrage in compensating the victims of social norm violations.

The finding that, controlling for empathic concern, moral outrage mediated the effect of the video manipulation on compensation further bolsters the claim of moral outrage's causal role.

We did not predict that compensation would be higher in the Bad Investment and Random Chance situations after watching the moral outrage-inducing video, which led us to conduct further tests to better understand those results. We observed that, although we chose the video due to its limited effect on other emotions, it also significantly affected empathic concern in the bad investment situation, which allowed for the possibility that it was the change in empathic concern, rather than moral outrage, which drove the effect. To test for this, we used mediation analysis, allowing for both empathic concern and moral outrage to serve as mediators of the video's effect on compensation in the bad investment situation. The finding that empathic concern, and not moral outrage, mediated the effect of the video on compensation is consistent with empathic concern, rather than moral outrage, driving compensation in the bad investment situation.

Similarly, we ran a mediation analysis in the random chance situation, testing the degree to which moral outrage mediated the effect of the video on compensation. Similar to the bad investment situation, we did not find moral outrage to be a significant mediator. Here we see a parallel of Study 1, where we found support for moral outrage *not* being a determining factor of compensation in the random chance situation, but this data does not speak to what may actually be the emotional determinants.

Study 2.b

Study 2.b closely mirrors the design of study 2.a, but focuses on the role of state empathic concern rather than moral outrage. In this study, empathic concern towards the person who lost their money was manipulated by having the participant either write a response to a prompt asking them to take the perspective of the person who lost their money, or to neutrally describe the interaction. Each respondent then participated in one of the three situations described in Study 1. We predicted that, consistent with previous work, those who responded to the high empathic concern prompt would compensate more in the situations not involving a norm violation, but that this pattern would not be present in the norm violation situation. Additionally, we predicted that, controlling for moral outrage as a covariate, empathic concern would mediate the effect of the perspective taking manipulation on compensation in the non-norm violation situations.

Method

We recruited 998 participants (472 men, mean age of 34) from the AMT platform to participate in this study. The design of study 2.b closely mirrored that of 2.a, with the key difference being our manipulation of empathic concern rather than moral outrage. Whereas anger is experienced as a general emotional state, empathic concern is, by its very nature, expressing concern *for* a particular person, which did not allow us to use a video manipulation. Instead, after reading the rules to the interaction, being assigned to their role as the third party, and seeing that the investor did not receive any money back, we had participants write in response to one of two prompts. In the control conditions, we asked participants to "objectively describe what has happened in the interaction so far". In the high empathic concern conditions, we asked participants to "describe the feelings and emotions Person A may be feeling right now".

Results

769 (77%) of participants correctly responded to 9 of the 10 comprehension questions. To ensure data quality, we only analyzed the responses from these participants.

Using the empathic concern scale as a manipulation check, we found that the perspective taking prompt lead to higher empathic concern relative to the control in the norm violation situation $t(269) = 2.92$, $p = .004$, the investment situation $t(244) = 3.34$, $p = .001$, and the random chance situation $t(250) = 3.40$, $p = .001$. These effects were moderate in size ($d = .36$ to $d = .43$).

We report mean levels of compensation in Fig. 3. Those in the bad investment situation who received the empathic concern prompt (M = 4.03) compensated significantly more than those who received the objective prompt (M = 3.05). In the bad investment situation, controlling for moral outrage as a covariate, participants' level of empathic concern significantly mediated the effect of the prompt manipulation on compensation $D = .45$, 95% CI = .1634 to .7010.

In the random chance situation, we did not find that those who responded to the empathic concern prompt (M = 3.13) compensated significantly more than those who responded to the objective prompt (M = 2.89) $t(250) = .822$, $p = .41$. Similarly, in the norm violation situation, we found no significant difference in compensation between

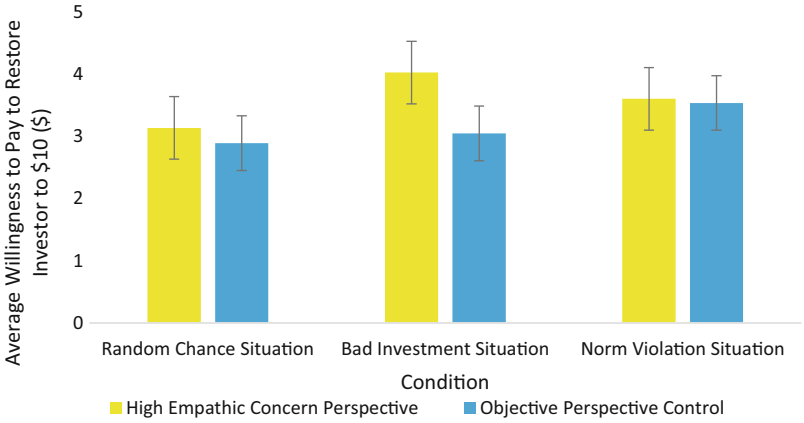


Fig. 3 Average willingness to pay of participants to restore investor to $10 by condition

those who received the empathic concern prompt (M = 3.61) and those who received the objective prompt (M = 3.54) $t(269) = .132$, $p = .90$.

Discussion

The empathic concern prompt leading to higher compensation in the bad investment situation supports the hypothesis and results from previous studies that empathic concern can drive compensation behavior. Mediation analysis further buttresses this finding, showing that empathic concern mediates the effect of the written prompt on compensation.

Consistent with our findings in Study 1, we did not find a significant effect of empathic concern on compensation in the random chance situation. This provides additional motivation for further investigation into what may be driving compensation in this context. We also do not find a significant effect of empathic concern on compensation in the norm violation situation, consistent with our general hypothesis that moral outrage, rather than empathic concern, drives compensation in the context of norm violations.

Study 3

We designed Study 3 (Thulin and Bicchieri 2017) to replicate and generalize the finding of Studies 1 and 2.a that, in the case of a social norm violation, moral outrage correlated with (Study 1) and drove (Study 2.a) participants' willingness to compensate. This study had two manipulations. First, participants were assigned to either the norm violation or bad investment situations previously described in Study 1. Second, participants were assigned to either a high moral outrage or neutral emotional video manipulation described in Study 2.a. We made two other key modifications from Study 2.a. First, participants interacted with each other for actual money rather than responding to hypothetical situations. Second, participant feedback suggested that the willingness to compensate measure used in Studies 1 and 2 was complex and therefore difficult to understand. We therefore substituted a simple transfer with multiplier as the dependent measure to improve participant comprehension. We predicted that those who watched the moral outrage-inducing video would compensate more than those who did not, but only in the social norm violation situation.

Method

We recruited 502 participants (243 men, mean age of 39) from the AMT platform to participate in this study. Participants were divided into two phases. Participants in Phase 1 read a description of a trust game, similar to those used in the previous studies. In the norm violation situation, investors were endowed with 50c and trustees with 0c. The investor could choose to either keep their 50c or transfer it to the trustee. If transferred, the experimenter tripled the amount to 150c. The trustee then had the option of whether to keep the entire 150c or to return 75c to the investor. As in the previous studies, the bad investment situation mirrors the norm violation situation, aside from one variation. Instead of the trustee *choosing* whether or not to return half the endowment, a randomizing device selected, returning half the endowment 80% of the time and none of the endowment 20% of the time. All participants were informed that the choices of future participants may impact their payoffs, but were not told in what way. Phase 1 was run until, for both the norm violation situation and the random chance situation, an investor chose to transfer their endowment to the trustee and the trustee chose not to return the sum. These final pairs were used as the focal dyads.

After establishing the focal dyads, all further participants were assigned to Phase 2. Each participant in Phase 2 read a description of the trust game outlined above. Participants were told that they were assigned to the role of a third party for an investor and trustee pair and given an endowment of 75c. They were told that if the investor chose to transfer their 50c to the trustee but 75c was not returned from the trustee to the investor, they would have the opportunity to transfer any amount of their 75c to the investor, and that the amount they chose to transfer would be doubled by the experimenter.

After reading these interaction instructions, Phase 2 participants were shown one of the two videos used in Study 2.a, to either induce moral outrage or serve as a neutral. After watching the video induction, participants were shown the result of one of the focal dyads, in which the investor chose to transfer to the trustee and either a randomizing device or the trustee selected not to return half the endowment, depending on condition.[1] After seeing the result, participants then chose how much of their endowment to transfer to the investor, which was doubled by the experimenter.

After making their selections, participants responded to the state moral outrage and empathic concern scales used in Studies 2.a and 2.b. Participants were immediately paid their 50c show up fee, and then paid their bonus amounts five to seven days later.

Results

385 (77%) of participants recruited for Phase 2 correctly responded to 4 of the 5 comprehension questions. To ensure data quality, we only analyzed the responses from these participants.

Using the four item state moral outrage scale as a manipulation check, we observed that the moral anger video significantly increased the level of moral outrage relative to the control video $t(486) = 4.87$, $p < .001$. This is a moderately sized effect $d = .44$.

The mean compensation values across conditions are shown in Fig. 4. We predicted that, in the norm violation situation, those participants who watched the moral outrage-inducing video would compensate to a greater amount. However, the difference observed was small and non-significant $t(196) = .40$, $p = .687$.

Due to this surprising result, we also investigated the partial correlations between moral outrage and compensation, controlling for empathic

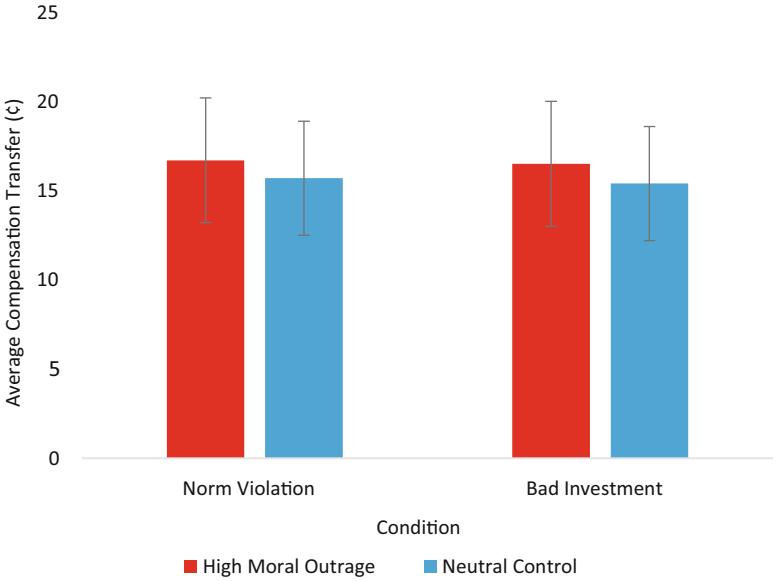


Fig. 4 Average amount transferred from participants to investor by condition

concern, in both the norm violation and bad investment contexts. Controlling for empathic concern, we found that moral outrage was significantly correlated with compensation in the norm violation context $r(197) = .15$, $p = .03$. This differed from the bad investment context, in which we did not find a significant relationship between moral outrage and compensation, controlling for empathic concern $r(184) = .09$, $p = .24$.

Discussion

The lack of an effect of the video manipulation on compensation was surprising, and inconsistent with the results of Studies 1 and 2.a. There were two key differences between the previous studies and Study 3, which may have affected the result. The first, and most concerning, is that the effect exists for hypothetical exchanges but does not generalize to exchanges involving actual money. There is reason to be suspicious of this possibility, as past research has shown that subjects drawn from AMT

respond similarly to hypothetical games as they do to those involving actual money, including in the specific context of trust games (Amir et al. 2012).

A second difference between the previous studies and Study 3 was the elicitation and measurement of compensation. In the previous studies, we used a hypothetical incentive-compatible elicitation of *the most one was willing to pay* to restore the investor to their original endowment. We gave each participant a series of binary choices, asking if they would be willing to pay X in order to restore the investor to $10, where X ranged from $1 to the third party's entire endowment of $10. After making their choices, one of the ten choices was randomly selected and carried out (e.g., if the "Would you be willing to pay $3 of your $10 to make person A end with $10?" question was selected and the participant chose "Yes", then the participant would have $3 deducted from their endowment, and the investor would receive $10).

We chose this method because willingness to pay has a high degree of granularity as compared to a single choice (e.g., only asking would you pay $2.50 to restore the investor to $10). It also measures the implicit lowest compensation trade-off ratio that a participant sees as making the transfer worthwhile, which we find to be a compelling proxy for one's willingness to compensate. For example, being willing to transfer $4 but not $5 to restore the investor to $10 implies the minimum acceptable compensation trade-off ratio between 2.5 and 2. This is in contrast to choosing an amount to transfer with a fixed multiplier, which has a much more ambiguous interpretation. As opposed to the willingness to pay measure, one cannot impute the minimum acceptable multiplier (as the multiplier is held constant). Instead, the amount transferred could indicate the amount the participant thinks the investor deserves, rather than how much one care that they are compensated.

Although the willingness to compensate measure had these desirable properties, feedback from participants in Studies 1 and 2 suggested that the method was very difficult to understand, and at a minimum, cognitively taxing. As we were particularly interested in the emotional determinants of the compensation decision, we chose to simplify the compensation measure in Study 3 by simply asking how much of their endowments participants wished to transfer to the investor, with a 2× multiplier. The lack of an effect in this case may therefore be because, while previous dependent measures assessed how badly the third party wanted the investor to be compensated, the current measure may be assessing the *amount* a participant believes the investor deserves, which may be less subject to moral outrage.

These concerns are partially assuaged by correlations within the data being consistent with Studies 1 and 2.a. Namely, the finding that moral outrage correlated with compensation in the norm violation situation but not in the bad investment situation, controlling for empathic concern, is the same pattern observed in Studies 1 and 2.a. This is consistent with the general hypothesis that moral outrage makes a significant unique contribution to the compensation of norm violation victims.

General Discussion

Taken together, these studies begin to reveal a richer landscape of emotional determinants of victim compensation than was previously identified. Studies 1 and 2.a found that on both the trait and state levels, moral outrage was associated with a willingness to compensate victims of social norm violations beyond the effect of empathic concern. In fact, when we controlled for moral outrage, or directly manipulated empathic concern, the data revealed no significant effect on the compensation of victims of social norm violations.

Also as predicted, the effect of moral outrage on compensation appears to be domain specific. We found no significant relationship between a propensity to feel moral outrage and willingness to compensate when a loss was due to random chance or a bad investment in Study 1. Despite finding significant differences in willingness to compensate in both the random chance and bad investment situations in Study 2.a, we found that moral outrage mediated neither of these effects. This result was consistent with Study 1, suggesting that moral outrage was not involved in driving compensation in these contexts.

Our finding in Study 3 that increasing moral outrage did not increase compensation in the norm violation context is in contrast with the pattern of results in Studies 1 and 2.a. One possible explanation for this discrepancy was the change in dependent measure. Whereas Studies 1 and 2 measured the most one is willing to pay to restore the investor to $10 (effectively measuring the lowest compensation trade-off ratio the third party is willing to accept), Study 3 measured the amount the third party chooses to transfer. The latter is at least partially determined by the amount a participant feels is the *correct* amount to transfer rather than just the degree to which they want to compensate the victim, which may be less influenced by moral outrage. Follow-up work may shed more light on this issue, which could have ramifications for the measurement of compensation in the future. Despite

this inconsistent finding, even in this study we found that moral outrage, controlling for empathic concern, correlated with compensation in the norm violation situation but not the bad investment situation, consistent with our previous pattern of results.

An unexpected but interesting result emerged when evaluating the relationship between empathic concern and compensation in the random chance situation. In the bad investment situation in both Studies 1 and 2.b, we found relationships between empathic concern and willingness to compensate. However, in the random chance situation, we found no such relationships between compensation and empathic concern. Further work is required to understand the distinguishing features between these two cases, and what other emotional determinants may be driving compensation when losses are due to random chance.

At first glance this general pattern of findings seems inconsistent with previous work demonstrating a relationship between empathic concern and third-party compensation of those who receive low offers in a dictator game (Leliveld et al. 2012). However, there are two possible ways to reconcile these findings. First, as reported in Study 1, there is a high correlation between moral outrage and empathic concern, which points to the importance of controlling for one in order to understand the influence of the other. As this previous work did not include such controls, it is possible that moral outrage, as a latent third variable, may account for the results. Second, other work has shown that people do not have strong personal beliefs of what divisions one *should* make in the dictator game, which is critical for the existence of a social norm (Bicchieri 2006). As no norm may exist in the dictator game situation, and therefore none may be violated, it would be reasonable for empathic concern, rather than moral outrage, to motivate third parties to compensate.

To conclude, much of Hardin's work focused on the problem of explaining the possibility of social order in large, complex societies. In small villages and large societies alike, social norms play a fundamental role in maintaining social order. The maintenance of social order in more complex societies may be addressed by analyzing the role that emotions play in supporting pro-social behavior (Elster 1999). Emotions like anger, or moral outrage, are felt (and acted on) in any environment where an established norm is violated. Additionally, the same emotion leads us to punish perpetrators and compensate their victims. In environments where interactions may be anonymous or among strangers, and monitoring behavior is difficult and costly, it seems that emotions play a large part in guiding

the sanctioning and compensatory behavior that maintains social order. Common expectations guide our behavior and determine our emotional reactions. Such emotions provide a common basis for pro-social behavior across the diverse communities that constitute a complex society.

Note

1. The method of using a focal dyad for all future decision has previously been used to maintain non-deception, as nothing false is told to participants, while increasing the efficiency of the study by minimizing the number of subjects necessary to achieve adequate power.

References

Amir, O., D.G. Rand, and Y.K. Gal. 2012. Economic Games on the Internet: The Effect of $1 Stakes. *PloS One* 7 (2): e31461.

Batson, C.D. 1981. Is Empathic Emotion a Source of Altruistic Motivation? *Journal of Personality and Social Psychology* 40 (2): 290–302.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*, 137–139. New York: Cambridge University Press.

Buhrmester, M., T. Kwang, and S. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6 (1): 3–5.

Carlsmith, K.M., J.M. Darley, and P.H. Robinson. 2002. Why Do We Punish?: Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology* 83 (2): 284–299.

Charness, G., R. Cobo-Reyes, and N. Jimenez. 2008. An Investment Game with Third-Party Intervention. *Journal of Economic Behavior & Organization* 68 (1): 18–28.

Chavez, A.K., and C. Bicchieri. 2013. Third-Party Sanctioning and Compensation Behavior: Findings from the Ultimatum Game. *Journal of Economic Psychology* 39: 268–277.

Coke, J.S., C.D. Batson, and K. McDavis. 1978. Empathic Mediation of Helping: A Two-Stage Model. *Journal of Personality and Social Psychology* 36 (7): 752–766.

Davis, M.H. 1983. Measuring Individual-Differences in Empathy – Evidence for a Multidimensional Approach. *Journal of Personality and Social Psychology* 44 (1): 113–126.

Elster, J. 1999. *Alchemies of the Mind*. Cambridge: Cambridge University Press.

Fehr, E., and U. Fischbacher. 2004. Third-Party Punishment and Social Norms. *Evolution and Human Behavior* 25 (2): 63–87.

Fehr, E., and S. Gachter. 2004. Altruistic Punishment in Humans. *Nature* 415: 137–140.

Fessler, D., and K. Haley. (2003). The Strategy of Affect: Emotions in Human Cooperation. In *The Genetic and Cultural Evolution of Cooperation*, ed. P. Hammerstein, 7–36. Dahlem Workshop Report. Cambridge, MA: MIT Press.

Fetchenhauer, Detlef, and David Dunning. 2009. Do People Trust Too Much or Too Little? *Journal of Economic Psychology* 30 (3): 263–276.

Gross, J.J., and R.W. Levenson. 1995. Emotion Elicitation Using Films. *Cognition and Emotion* 9 (1): 87–108.

Haidt, J. 2003. The Moral Emotions. In *Handbook of Affective Sciences*, ed. R.J. Davidson, K.R. Scherer, and H.H. Goldsmith. Oxford: Oxford University Press.

Hardin, R. 1999. From Bodo Ethics to Distributive Justice. *Ethical Theory and Moral Practice* 2 (4): 399–413.

Leliveld, G.J., E. van Dijk, and I. van Beest. 2012. Punishing and Compensating Others at Your Own Expense: The Role of Empathic Concern on Reactions to Distributive Injustice. *European Journal of Social Psychology* 42: 135–140.

Lerner, J.S., J.H. Goldberg, and P.E. Tetlock. 1998. Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility. *Personality and Social Psychology Bulletin* 24 (6): 563–574.

Nelson, M., and M. Zeelenberg. 2009. Moral Emotions as Determinants of Third-Party Punishment: Anger, Guilt, and the Functions of Altruistic Sanctions. *Judgment and Decision Making* 7 (7): 543–553.

Piazza, J., P.S. Russell, and P. Sousa. 2013. Moral Emotions and the Envisaging of Mitigating Circumstances for Wrongdoing. *Cognition and Emotion* 27 (4): 707–722.

Pillutla, M.M., and J.K. Murnighan. 1996. Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers. *Organizational Behavior and Human Decision Processes* 68 (3): 208–224.

Preacher, K.J., and A.F. Hayes. 2008. Asymptotic and Resampling Strategies for Assessing and Comparing Indirect Effects in Multiple Mediator Models. *Behavior Research Methods* 40 (3): 879–891.

Schachter, S., and J. Singer. 1962. Cognitive, Social, and Physiological Determinants of Emotional State. *Psychological Review* 69: 379–399.

Thulin, E., and C. Bicchieri. 2015. I'm So Angry I Could Help You: Moral Outrage as a Driver of Victim Compensation. *Social Philosophy and Policy* 32 (2).

———. 2017. Anger or Empathy: What Drives Victims' Compensation? BeLab Discussion Paper, University of Pennsylvania, May.

Toi, M., and C.D. Batson. 1982. More Evidence That Empathy Is a Source of Altruistic Motivation. *Journal of Personality and Social Psychology* 43 (2): 281–292.

Wakslak, Cheryl, J., J.T. Jost, Tom R. Tyler, and E.S. Chen. 2007. Moral Outrage Mediates the Dampening Effect of System Justification on Support for Redistributive Social Policies. *Psychological Science* 18 (3): 281–292.

Wenzel, M., and I. Thielmann. 2006. Why We Punish in the Name of Justice: Just Deserts Versus Value Restoration and the Role of Social Identity. *Social Justice Research* 19 (4): 450–470.

Index

Note: Page number followed by 'n' refers to notes.

T. Christiano et al. (eds.), *Morality, Governance, and Social Institutions*, DOI 10.1007/978-3-319-61070-2