Para-Ideal Theory and the Strategic Justification of Democracy[1]

For presentation at the workshop on Non-ideal and Institutional Theory at the Canadian Political Science Association annual meeting, June 2, 2010.

Paul Gowder
Stanford University
Department of Political Science
616 Serra St., Encina Hall West, Room 100
Stanford, CA 94305-6044, USA

pgowder@stanford.edu

(*Working draft: as is standard in such situations, please don't cite or quote, and please forgive glitches/omissions/missing citations, etc.*)
[Note to fellow conference participants: I would particularly appreciate feedback on the final section, pertaining to democracy — is it even worth keeping, or is it simply over-ambitious?]

Introduction

In political philosophy and theory, we can often describe moral obligations from individual and from collective standpoints. Consider the problem of distributive justice.[2] From the collective standpoint, we often say that the community as a whole has an obligation to bring about a just distribution of resources (or welfare, or primary goods, etc.). We also speak of the same moral concern from the individual standpoint, saying that individuals have an obligation to, *inter alia*, support just institutions, pay their taxes, contribute to the less well-off, etc. Or suppose one's government is waging an unjust war. We again speak of the obligation to put a stop to the war as collective and the obligation to work toward doing so as individual, and we often think that individual members of the community are (to at least some extent) blameworthy for the unjust war just as is the community as a whole, and even if those individuals do not directly participate in that war.[3] This dual perspective scales up to higher levels of collectivity:

---

[1]A very early (and vastly more flawed) iteration of this project was presented at the 2009 Pavia Graduate Conference in Political Philosophy. An iteration with different flaws was presented at the conference on Democracy as Idea and Practice, University of Oslo, 2010. I thank the participants at those conferences for copious helpful feedback, particularly Francesca Pasquali, Federico Zuolo, and Lene Bomann-Larsen. I also thank Nicole Hassoun, Ruth Kricheli, Peter Northup, and Daniel David Slate for very helpful comments on various versions.

[2] I will use the term "moral concern" to denote some morally relevant problem, like distributive justice, about which individuals and collectives may have duties. Distributive justice will supply the running examples throughout this paper, but the issues identified here are not limited to that moral concern. They may arise whenever coordinated individual action is necessary to achieve some kind of obligation in political morality, such as, for example, to end political violence or to build legitimate institutions.

[3] This sort of reasoning, for example, underlies much of the argument for imposing reparation obligations on otherwise-innocent individuals who were members of states that perpetrated great injustices [cites: Satz, AP].

when considering global justice, we can speak of the duties of individual states as well as the duties of the global community of states as a whole.

The relationship of moral concerns viewed from the collective standpoint and from the individual standpoint is not terribly well-studied, but we seem to share one major intuition: we ordinarily expect it to be the case that if every individual in a community is satisfying his obligations with respect to a given moral concern, then the community as a whole is satisfying its obligations under the collective description.[4] This assumption is ordinarily (but not always) tacit, and can be seen in our discussion of ideal theory. The conventional understanding of "ideal theory" in political philosophy is that it is the theory of (one's duties in) those situations in which everyone (else) is complying with his moral duties. The assumption at issue can be seen in the fact that ideal theorists feel themselves entitled to *stop* there — no ideal theorist asks the further question "given that everyone is complying with his political-moral duties as specified by my normative theory, how do we bring it about that the community is complying with *its* duties? Several explicitly describe the full compliance situation (sometimes with, sometimes without, a loosely-specified caveat about "favorable circumstances") as equivalent to a fully just, moral, etc. political community.[5] Put differently, it is ordinarily assumed that, in the full-compliance situations described by ideal theory, no moral improvement is possible (or necessary[6]) from the collective standpoint.

In section I, I refute that assumption. I demonstrate that there can be moral circumstances analogous to sub-optimal coordination equilibria in game theory—in which each individual is fully complying with her moral obligations, but, if the community as a whole could modify all individuals' behavior en masse, it could achieve morally superior collective behavior, and no individual can unilaterally modify her behavior except at a moral cost.[7] I call these *para-ideal* circumstances, since they have one property of ideal theory, traditionally conceived (everyone is complying with his moral obligations), but not the other (the state of affairs is less than, well, *ideal*).

In section II, I argue that the state's coercive power can solve para-ideal situations. I then argue that there is a special role for democratic institutions in this process. Democracies permit each citizen, motivated by moral ends rather than coercion, to vote for the coercive pursuit of a

---

[4] This, of course, would also entail the contrapositive: that if the community as a whole is not satisfying its obligation, then at least one individual is not satisfying his obligation. It does not, however, entail that if the community as a whole *is* satisfying its obligation then all individuals are satisfying their obligations — and that claim is neither commonly assumed nor even minimally plausible. Consider, for example, a democracy where every citizen but one complies with her duties to vote for a just order. The community as a whole achieves justice, because it wins at the ballot box, but without the participation of that individual.

[5] For example, Rawls suggests that principles of justice formed under the presumption of full compliance "defin[e] a fully just society, given favorable conditions." (John Rawls. 1999. *A Theory of Justice* (rev. ed.). Cambridge: Harvard University Press. 308-9.) Zofia Stemplowska, "What's Ideal About Ideal Theory?" *Social Theory and Practice* 34:319-340 (2008), pg. 332, equates full-compliance theories with "a final landmark of where we ought to be."

[6] Frederico Zulo has aptly pointed out to me that there could be undemanding versions of ideal theory which recognize the possibility of moral improvement, but do not require it (i.e., collective superogation).

[7] By "moral cost," I mean that an individual who unilaterally modifies his behaivior will be behaving worse, from the moral standpoint, after the modification.

collective moral end, and thus makes it possible for each citizen to make her actual pursuit of that end conditional on each other citizen's pursuit of that end. By contrast, in a non-democratic state, coordination can be achieved by the command of a leviathan, but that coordination does not recruit the moral motivations or expressions of individuals. This is a moral advantage that democratic governments have over all other forms of social organization, which I will call *the strategic justification of democracy*.

I. Moral Coordination Problems and Game Theory

Suppose that everyone has individual duties with respect to some moral concern. These duties might call for unilateral behavior[8], and/or for the individual use[9] or shaping[10] of collective institutions (e.g., by political activism), but they still are, *ex hypothesi*, *individual* duties. Suppose further that the community as a whole has a duty with respect to that moral concern (e.g., to see that its poor get fed), and that the individual duties of each citizen are at least in part duties to do their part in bringing about the satisfaction of the collective duty (which we may assume is practically achievable). Finally, suppose that the morally relevant behavior of the community as a whole can be fully described by describing the individual behavior of each of its members (that is, the community's behavior is *reducible* to the behavior of its members, and nothing else). Intuitively, we would think that full satisfaction of each individual duty would necessarily lead to the full satisfaction of the community's duty. It is that intuition that I propose to refute in this section.

A. The Strategic Nature of Collective Morality

Many of the moral concerns with which political philosophers are concerned are concerns that can only be addressed with coordinated action from multiple people. For example, no individual can achieve distributive justice in Rawls's sense, since justice is a property of the institutions that make up the basic structure of the society, comprising the coordinated behavior of masses of people. No individual can even pursue Rawlsian distributive justice except by acting through and on the institutions of a political state. Viewed from the perspective of the individual concerned with satisfying his duty to advance the ends directed by those concerns, such an individual's duties are essentially *strategic*. By "strategic," I mean that our abilities to individually pursue the satisfaction of collective obligations are interdependent: my ability to fulfill them depends on your actions, and vice versa. An action that is efficacious at bringing about the moral end given one pattern of behavior from other people may be completely

---

[8] Such as giving to the poor.

[9] Peter Singer, for example, points out that individual duties of global distributive justice may best be satisfied by the use of collective institutions [cite: famine, affluence, morality]

[10] For example, for Rawls justice is a property of basic structures. But a Rawlsian might plausibly think that individuals in non-ideal situations have a duty of justice to bring it about that just basic structures are implemented, and in ideal situations a duty to support and maintain those structures at least to the extent of, for example, not disobeying redistributive tax laws.

inefficacious or even detrimental with another pattern of others' behavior.[11]

Since “ought implies can,” the answer to the question “is A obliged to X” depends on the extent to which A may act to X. When X is "bring about some property of a political community," and A is an individual, the extent to which A is capable of X-ing, and thus obliged to X at all, depends on the behaviors of the other members of that community. This relationship is symmetrical and circular: *mutatis mutandis*, the obligations of the others depend on A's behavior, and (at least in full-compliance situations), that behavior itself depends on A's moral obligations. A's behavior depends on A's obligation, which depends on B's behavior, which depends on B's obligation, which depends on A's behavior. Put differently, A’s *moral obligation itself,* in such circumstances, is strategic. This is already partially recognized by those political philosophers who acknowledge that moral rules should change in non-ideal circumstances, and that general moral propositions may be false insofar as they are stated without regard to their consequences under conditions of partial compliance.[12]

We can sum up the immediate implications of the last few claims in one sentence, which I will label *the improvement condition* (IC). IC: If A cannot make a moral improvement in the circumstances in which she finds herself by unilaterally changing her behavior[13], A is fully complying with her moral duties.[14]

Concepts from game theory will help make sense of the strategic facets of these obligations. I will not argue that game theoretic proofs are directly applicable to these problems (though they may be), for it is not yet established that moral choices are maximization problems over anything sufficiently analogous to the cardinal rankings of von Neumann-Morgenstern utility to simply import the mathematics. However, the strategic reasoning underlying game theory *does* directly apply.

---

[11] For example, I may satisfy my duty to help relieve famine in a poor country by sending money to that country's benevolent government. But if that govenment becomes malevolent and starts to use its money to buy weapons for oppression, an identical contribution becomes counterproductive and arguably even blameworthy.

[12] Perhaps the quintessential case is Murphy's "compliance condition" (in chapter 5 of Liam Murphy. 2000. *Moral Demands in Nonideal Theory*. New York: Oxford University Press.), which represents a proposal for a general bound on moral principles to take into account the possibility of partial compliance.

[13] Under the head of "unilateral change of behavior" we should include interactive behaviors like attempting to convince others to behave dutifully.

[14] Naturally, this elides a number of controversial questions about the sense of possibility in play, e.g., whether we might say that someone is complying with her duties when she is merely psychologically unable to carry them out. These questions are not in play in this paper, for my theory here is limited to the sorts of obligations that are directed at bringing about a collective end, and the sense of possibility at issue is the possibility of achieving such an end without the cooperation of necessary others. I take it that it will be totally uncontroversial that Peter is not obliged to achieve or pursue an end that requires Paul's help when Paul is unwilling to do so.

If there are no supererogatory duties, it also seems true that if an individual *can* make a moral improvement, she is obliged to do so, but this claim is not relevant to this paper. Incidentally, I assume throughout this paper that there are no supererogatory duties at stake. By doing so, I do not purport to be taking a position on the debates as to whether such duties exist, or whether there may be any such duties in the political sphere. Rather, I submit that they are simply not presently relevant for any of the urgent problems facing political actors as well as political theorists: we have more than enough trouble meeting even a small fraction of the mandatory duties of political morality that every serious theorist thinks we have.

Assuming that community members have some moderately well-behaved moral aims (aims that can generate an ordinal ranking over the various behaviors that oneself and one's community can engage in, from most to least morally acceptable), the essential logic of game theoretic reasoning can be used to capture the ways in which the moral ends that one citizen aims at are constrained by the moral ends that other citizens aim at. For example, we can think of a *moral Nash equilibrium* as a situation in which a group of citizens is behaving such that no citizen can achieve a result that he thinks is morally superior by a unilateral action.

Consider a more concrete example of the usefulness of this tool. Bernard Williams has provocatively argued that political communities sometimes need citizens in positions of power who are willing to swallow their moral compunctions for a greater good.[15] Suppose that Williams is right, and that sometimes politicians are obliged to act immorally in this special (and somewhat odd) sense. Now suppose a community is faced with the choice about whether or not to start a necessary, but deeply regrettable, war. The 101 member senate, in whose hands the decision is placed, currently has 49 votes against the war, and 50 votes in favor. Either of the two remaining senators may be the deciding vote in favor of starting the war.

Assuming that the senate as a whole fails in its duty to the community if it does not bring about the war, the two remaining senators face a problem: who is to get the blood on her hands by voting for the war? We can understand this problem one of two ways. First, we might think there is a genuine moral conflict: each senator has an obligation not to dirty her hands by voting for this nasty war, but also has an obligation not to stand in the way of the war. Alternatively, we might think that the senators are acting under moral uncertainty: it seems to each that it's more likely than not that the war is justified, but it would be better, from the standpoint of the conscience of each, if the war happens without her active participation.

Given either interpretation of this scenario, the best moral option for each is to not stand in the way of there being a war, but without dirtying her hands by voting for it. Second-best is to vote for the war and for there to be a war, and third-best is to not vote for a war and for there to be no war. Examining the problem strategically, we see that we have a the moral equivalent of the standard battle of the sexes game with two pure strategy Nash equilibria: 1) A votes yes, B votes no; and 2) A votes no, B votes yes. Any other paring of votes means that someone can make a moral improvement by switching her vote. The game-theoretic approach allows us to see that the senate as a whole has a moral burden to allocate. The group will satisfy its moral obligations regardless of which senator dirties her hands, but it will not do so unless one does so.[16] This demonstrates the sort of insight that looking at collective morality strategically can offer.

---

[15] *Moral Luck*, Cambridge University Press, 1981, ch. 4.

[16] A similar sort of problem is faced by the members of a firing squad, who may believe that it is *probably* right for the execution to proceed. The tradition (or urban myth) of giving some members of such a squad blank ammunition can be seen as a way of permitting the individual shooters to satisfy their consciences without compromising the collective end: the prisoner is still executed, but each member of the firing squad is slightly less likely to have dirtied his hands in the course. This suggest an analogy between that institution and mixed-strategy equilibria in conventional game theory, which is, alas beyond the scope of this paper.

B. Para-Ideal Theory and Moral Public Goods

The strategic approach to collective morality also reveals that it is possible to have a state of affairs in which everyone is satisfying his individual moral obligations to contribute to a collective moral end, yet collective moral improvement is possible when citizens find themselves in sub-optimal moral Nash equilibria. I will illustrate with another example.

Suppose that a community has an obligation to tend for its destitute members. It may do so either by building a homeless shelter or by building a soup kitchen. Building a soup kitchen would be better than doing nothing, but not nearly as good as building a homeless shelter, which, let's say, fully discharges that community's duty. Each individual has a duty to contribute to the community's satisfaction of its obligation toward the poor. Assume that each individual is completely willing to comply with this duty, but is uncertain about the willingness of each other individual (which entails that there is no "common knowledge" in the traditionally strong game theoretic sense). Each individual has a $100 budget constraint. If everyone gives $100, they can build the shelter, but if any citizen gives only $50, they can only build the soup kitchen, and if they build the soup kitchen, any contributions above $50 will go to waste. Moreover, let us say that if a citizen only contributes $50, he will be able to achieve some other moral good with the $50 saved, though not one that is nearly enough to make up for the loss of the homeless shelter (perhaps he is able to take in a stray dog). Contributions are determined simultaneously.

For purposes of simplicity, we can assume that there is a very small population of players for this game — say, only two people (plus a third homeless person who is not a player insofar as he is unable to contribute). The point revealed by this example ought to be generalizable to more complicated n-player games that are less analytically tractable. In our example, there are two moral equilibria: 1) everyone donates $100 to the shelter (EQ-SHELTER) and 2) everyone donates $50 to the kitchen and $50 to a dog (EQ-DOG). In each case, nobody can make a moral improvement unilaterally. Should the community find itself in EQ-DOG, everyone can genuinely say that she is making the morally best individual choice, given the constraints imposed on what she may achieve by the behavior of others. Indeed, if one citizen thinks everyone else will be giving $50, she will incur a moral *cost* — will arguably be *blamable* — by giving $100 and wasting money that could feed a dog.

If "ideal theory" is simply the theory of full compliance, then EQ-DOG qualifies for the title. After all, IC entails that *no individual is blamable* for the amount of his contribution in EQ-DOG, so long as each individual's duty is understood to require doing her part for the best overall collective result that her behavior can achieve.

This might be controversial. A certain kind of Kantian might be attracted to the notion that each individual citizen has a duty to contribute $100, just because that contribution is what, when universalized, would fully satisfy the community's duty. And some consequentialists might go along with them, reasoning that the decision procedure for agents attempting to achieve morally optimal consequences can be separated from the truth conditions for moral claims, such that we can fairly say, qua truth claim, that no improvement is possible in the $50 equilibrium, but, *nonetheless*, each individual ought to contribute $100, regardless of what each other individual is doing, and that this will lead to the best consequences, namely, the achievement of the homeless shelter. (We can call this the "what if *everybody* said his vote didn't count?"

argument.)

But I think we must reject both of those moves. Each amounts to demanding that we simply ignore the strategic circumstances when judging the moral worth of an individual contribution to a collective end. That seems plausible only because we know that it reaches the right result (collectively) when we know (from the third-person omniscient narrator standpoint — our theorist's "view from nowhere") that each individual is properly motivated to comply with their duties. But it seems much less plausible when we consider things from the perspective of an agent who genuinely does not know how his fellows will behave. Compare this to a more familiar analogous problem, that of voting for a minority political candidate. Suppose a citizen believes that candidate A is the best choice, followed by candidate B, and then the disastrous candidate C. If she also believes that the other voters are roughly evenly split between candidates B and C, and that A has no hope of winning, can we genuinely advise her to choose A over B, even when, from her perspective, that amounts to bringing about all of the horrors that come with C's victory? Instead, we should, when evaluating the morality of an agent's actions, respect the epistemic constraints on that agent.[17]

This first-personal perspective represents a general boundary around my argumentative ambitions. I do not engage questions about whether an agent who brings about a morally inferior state of affairs due to incomplete information is in any way blameworthy for doing so from some third-personal perspective, or whether such an agent has done something impermissible even though not blameworthy and thus isn't *really* in a state of full compliance, and the like. These questions are beside the point. From the first-personal perspective of an agent who is genuinely concerned with doing the best moral action available to her and bringing about the morally best state of affairs possible, such an agent would rather be one who does the full-information right action instead of the one who does the partial-information right action, but would not blame herself for making the best she could out of the information she has. It is in this sense that I say that a society full of such agents has full compliance with individual moral obligations.

For all this, is is unsatisfactory to call the $50 equilibrium “ideal theory,” since there is a morally better course of action for the community. I propose we call the theory of situations in which communities are in sub-optimal moral equilibria *para-ideal theory*.

Para-ideal circumstances are those in which individuals are fully complying with their moral duties with respect to collective ends, but moral improvement is possible for the

---

[17] Moreover, it might not be the case, even from a third-person omniscient standpoint, that every individual is properly motivated. If everyone but one citizen is *in fact* motivated only to give $50, we surely wouldn't say that the one properly-motivated citizen must give $100. (Or, were we to do so, we would be rightly accused of unreasonable rigorism of the sort that disregards the circumstances external agents find themselves in [citation: Tamar Schapiro]). But if that's true, then we can't demand it of a citizen who doesn't know how others are motivated either, unless we are willing to bite the heavy bullet of suggesting that an individual's moral duties can change based on facts that are completely unknown to him.

To clarify, we may say that an individual in such a circumstance has *moral reason* to give $100, in virtue of the fact that everyone is properly motivated such that everyone ought to give $100 in the abstract sense of "ought" that does not regard feasibility constraints (a sense of "ought" I borrow from Andrew Mason. 2004. "Just Constraints." *British Journal of Political Science* 34:251-268, 257.). But in the concrete sense of ought that does depend on "can," sush an individual is unable to follow that moral reason, and thus not obliged to do it — she is fully compliant with all the moral duties with which she has the power to comply.

community as a whole.[18] They are distinguished from ideal circumstances by the possibility of collective improvement (community full compliance), and from non-ideal circumstances by the existence of individual full compliance.

	Full individual compliance	Partial/no individual compliance
Collective improvement impossible	Ideal theory	(irrelevant to this paper)
Collective improvement possible	Para-ideal theory	Non-ideal theory

The key feature of para-ideal circumstances is *coordination failure.* Individuals are uncertain about the behavior of one another, such that even if they all are properly motivated to fulfill their individual duties, they may be unable to achieve the community's collective duty. With that understanding, we can see that our political world is full of potentially para-ideal circumstances. I will highlight the features of the running homeless shelter/soup kitchen example that make the situation ripe for coordination failure even when all agents are correctly motivated:

1) There are several possible collective outcomes from aggregating individual behavior in a community, and those outcomes can be ordered by their moral value;

2) The group must act in a coordinated fashion in order to achieve the best collective moral outcome, otherwise they get an inferior result;

3) At least some individuals are uncertain what their fellows will choose to do, such that they might not be able to coordinate on the best result; and

4) There is a moral cost to pursuing the optimal collective end without achieving it, such that it is the best choice of those who think their fellows will pursue some sub-optimal end to themselves pursue that end rather than the best end.[19]

---

[18] Here, I am conflating for simplicity three ideas in the concept of full compliance: 1) everyone has the right moral motivation, 2) everyone knows the relevant moral facts (though not necessarily the empirical facts that bear on applying moral truths to specific situations), and 3) everyone actually takes the morally best action. It is not necessary for purposes of this paper to treat these elements separately, and I will sometimes use correct motivation to stand for all three.

[19] The shelter/soup kitchen example was carefully constructed to contain a stepwise function translating contributions into collective results, so that $100 contributions would be wasteful and morally costly in the $50 equilibrium. It might be objected that this poorly resembles reality. (I thank Adam Fraser for raising this point.) But there are many similar situations described by non-continuous functions. Consider again the problem of strategic voting: in a two-party-dominated state, a plurality of citizens might think some radical third-party candidate is best, but vote strategically to elect their second-best choice, a major party candidate, out of ignorance of the beliefs of other voters. This sort of stepwise coordination failure also comes in the form of "tipping point" problems, such as the one discussed in Lecture 2 of Glenn Loury's 2007 Tanner Lectures, in which individuals find themselves compelled to carry guns once they believe a sufficient number of their fellow community members are doing so.

The uncertainty feature deserves special attention. There are several ways in which individuals can be uncertain about each others' willingness to pursue collective moral ends. They can be uncertain about each others' *beliefs* — A may think that B disagrees about the correct result. Alternatively, they can be uncertain about each others' *motivations* — A may think that B knows the right thing to do, but is unwilling to do it, as when B is unwilling to make sacrifices for the collective good, or when B suffers from weakness of will. Both of these forms of uncertainty are analogous to the players having incomplete information about each others' preferences in the ordinary game theoretic context.

There is also a third type of uncertainty in multiple-equilibrium situations in which more than one equilibrium is acceptable, but only one may be selected. Suppose, for example, that the community has the resources to either build a hospital or send military aid to an oppressed neighbor. If individuals do not know one each others' choices, it could find itself in a total coordination failure where neither gets accomplished because the community divides its efforts. This prospect is particularly worrying when political morality does not fully determine the choice to be made, i.e., because the two options are on a par or incommensurable — in which case even a population of citizens who have common knowledge of their good will and moral beliefs may be unable to coordinate.

In any of these para-ideal circumstances, if everyone could coordinate a mass change in behavior, the community could make a moral improvement, but individuals cannot make unilateral improvements. The object of para-ideal theory is to permit properly motivated citizens to act morally, both indvidually and collectively. It does so by advising on the task of moving from a sub-optimal to an optimal (or closer to optimal) equilibrium—to allow individuals to collectively modify their actions in order to make otherwise impossible moral improvements available.

II. The Role of the State in Para-Ideal Circumstances

A. Leviathan as Moral Coordination Device

It is well-known that political states, in virtue of their coercive power, are particularly well suited to solve coordination problems of the non-moral sort. It's natural to look to the state to solve moral coordination problems as well. That instinct is correct.

Let us continue with the running example. Should the community find itself in EQ-DOG, they could solve the problem by assigning the decision and implementation of the soup kitchen/ shelter question to a state. Suppose an absolute ruler—Leviathan—exists in the community. Leviathan, with his coercive power, can give each citizen reason to believe that each other citizen will contribute $100, rather than $50 – because he is forcing everyone to do so. Since a citizen who believes that every other citizen will donate $100 makes the best moral choice by donating $100 herself, the state makes it possible to escape the para-ideal equilibrium by solving the information problem that creates it. By conferring knowledge on each citizen, Leviathan makes it such that it is unequivocally the best moral choice, from an individual perspective, to do the thing that brings about the correct moral result from the collective standpoint. And this can make-up a justification of the state, insofar as the existence of a central coercive authority is necessary to achieve collective moral ends in situations of para-ideal coordination failure. (Note

that this justification is still from the first-person perspective — that is, the claim is that morally motivated agents ought to want to have a state to bring about their moral ends.)

Call this the *moral-strategic justification of the state* (MSJ). It was perhaps foreshadowed by Locke's point about the need for a neutral judge. The Lockean state of nature can be interpreted as a para-ideal situation: everyone (contra Hobbes's story) might be perfectly willing to respect the property of others, however, because of uncertainty about one another's beliefs and motivations, they may be unable to coordinate on a single option from among the available systems of property rights. As a result, they cannot attain the moral public good of a unified property system, and this leads to conflict and insecurity as people acting in good faith enforce different property regimes. Enter the universally accepted neutral judge in the form of the state, who makes the decision and permits everyone else to coordinate.[20]

It is also worth briefly comparing this argument to that advanced in Kavka's final paper.[21] Kavka suggests that even perfectly virtuous "angels" could have disagreements about the moral ordering of various actions, such that they would require a government to bring about the best moral results. Kavka too offers a justification of the state based on the properties of interactions between a multitude of even good citizens, and I see nothing to disagree with in his approach. However, mine is different in that it centers not on disagreement about the practical implications of moral principles, but on uncertainty about one's fellow citizens' beliefs and behavior — the argument I advanced above elucidates a problem even for citizens who happen to have all the *same* beliefs and perfect motivations about moral truth and its practical implications, so long as they don't *know* that they have the same beliefs and perfect motivations. And while Kavka mentions in passing the possibility of coordination problems due to incomplete information, he gives no details. Yet, in view of the well-known human tendency to hastily attribute moral and motivational failings to others[22], finding themselves in para-ideal situations likely to be the most important way that properly motivated citizens could go awry in anarchy.

MSJ is distinct from, and in two respects superior to, the conventional approach that justifies the state on the grounds that it permits the coercive supply of *ordinary* public goods, which would not otherwise be provided.[23]

First, it is compatible with more optimistic views of humanity. Some object that the traditional public goods defense of the state imagines an excessively selfish and untrustworthy sort of human — that only an extreme *homo economicus* type would be unwilling to take the risk necessary to contribute to public goods in the absence of a coercive guarantee that others too

---

[20] In this situation, Leviathan's coercive power may not even be necessary to achieve coordination. If citizens know that everyone is acting in good faith, and simply disagree about what best system of property rights is, then Leviathan's announcement may create a focal point [cite Schelling on focal points — enough space to explain?] around which citizens coordinate voluntarily.

[21] Gregory S. Kavka. 1995. "Why Even Morally Perfect People Would Need Government." *Social Philosophy and Policy* 12:1-18.

[22] [citations: fundamental attribution error literature, demonizing political opponents/polarization literature, etc.]

[23] [citation to traditional justification — Friedman? — and a word or two of description]

would contribute.[24] MSJ, by contrast, relies on features of collective action that hold true even for an altruistic population.

Second, the traditional approach is open to the objection that no individual is necessarily morally obligated to bring about public goods.[25] The moral public goods approach is less beset by this problem, for each individual is, ex hypothesi, already morally obligated to work to bring about the moral end on the basis of which a coercive authority is justified.[26]

B. Democracy and Moral Motivation

An autocratic despotism can use compulsion to escape para-ideal circumstances. "Give $100 or you will be shot" is sufficient incentive for agents with no moral motivations whatsoever. But this is somewhat unsatisfying for just that reason: the collective moral ends that such a state instantiates do not express the moral motivations of its citizens.

By "express the moral motivations of its citizens," I mean that in the absence of coercion (whether from a state or from other sources of coercion, like overwhelming social pressure), an individual who contributes to a collective moral end communicates to others, at minimum, a willingness to participate in that end, and may express an actual normative endorsement of that end.

Moral expression serves several purposes. The first is informational. Let's modify our running example to be a repeated game: in order to maintain a homeless shelter, citizens must contribute $100 each month, while in order to maintain a soup kitchen, citizens must contribute $50 each month. Suppose that by good luck our citizens manage to coordinate on EQ-SHELTER in the first round of this game, even in the absence of a state. In subsequent rounds, citizens now have information about each others' dispositions: each has good reason to believe that other citizens are disposed to contribute $100 in subsequent rounds, and thus that the homeless shelter is sustainable. Thus, in subsequent rounds, it will continue to be the best moral choice of each to contribute $100 — EQ-SHELTER, once reached once, will tend to persist. By contrast, Leviathan's equilibrium will be fragile: people get no information about what their fellows would do in the absence of coercion from what they do under coercion. Thus, the Leviathan solution requires the *continuous* threat of coercion even if nobody's motivations change — a threat that is both objectionable and costly if it can be avoided.[27]

---

[24] This objection is raised, for example, by Sartwell [cite "against the state" passage bookmarked on Kindle]

[25] Jonathan Wolff has raised this objection against Nozick's story of the growth of private "protection agencies" into a state — he is unable to convincingly explain why these protection agencies are entitled to force non-subscribers who live within their "territory" to participate in the shared project of defense [citations to Wolff book on Nozick]

[26] My approach is not *completely* free of this problem, since "individuals are morally obliged to X" does not always entail that "individuals may be subject to state coercion to bring about X." But we're at least a little closer than Nozick was, since it is much more plausible that people may be coerced sometimes to bring about the satisfaction of their moral obligations than it is that they may be coerced to bring about something that is merely good for them and everyone else.

[27] One might object that this could happen in anarchy too, if people would simply talk to one another. This is true, but talk in an anarchy is cheaper than a vote to coerce oneself to act in the way to which one claims to be committed. Admittedly, such a vote may be fairly cheap in a democracy too, if the voter believes that other citizens might not vote likewise — but it's still a more significant commitment than mere words.

Second, many have argued that there is a distinct moral value to acting from moral motivations. An individual who has a genuine choice about participating in some collective moral end is given an opportunity to reflect on her values and beliefs, and if she does participate in that end, she does so out of moral conviction. By contrast, while an individual who lives in a state that coerces participation in collective moral ends *may* also act out of moral conviction—that is, her participation may be overdetermined, such that moral conviction and coercion are each sufficient to bring it about—she *need not* do so. Individuals who conserve cognitive resources may never bother to reflect and form convictions about the rightness of those acts that the state simply compels. In fact, coercion might actually weaken citizens' moral motivation to the extent they come to resent Leviathan's imposing participation on them, or learn to substitute Leviathan's judgment for their own.

This second problem may seem beside the main point of this paper — the inability of citizens to act autonomously under autocracy is a general objection to autocracy, that does not seem to specifically limit Leviathan's power to solve para-ideal situations.[28] But recall that the point of para-ideal theory is to permit citizens to act morally both individually and collectively. If the solution to the problem of achieving collective moral behavior is to undermine the individual moral behavior that existed beforehand, then we cannot say that the para-ideal circumstances have truly been solved.

For those reasons, simply using Leviathan to achieve moral coordination is less than fully satisfactory. However, a *democratic* state can actually recruit the moral motivations and expressions of individual agents in support of the escape from para-ideal circumstances.

Let us suppose, in our running example, the decision between coercive state implementation of the shelter and the soup kitchen is put to a vote. The one-round game becomes a two-round game: each citizen (with motivations and knowledge as before) chooses whether to vote for a shelter or a soup kitchen, then the state coerces each citizen to contribute the amount chosen by the electorate.

In this example, the voting stage is not strategic, in the following sense: the morally relevant consequences of a citizen's vote choice do not change depending on what other citizens do — in game theoretic terms, a vote for the homeless shelter is each citizen's (morally) best response regardless of what the other citizens do; it is a strictly dominant strategy. This is because a vote for the shelter (unlike the $100 donation in the unmodified example) does not come with a (moral as well as practical) cost if others do not go along: nothing is wasted if it happens that every other citizen happens to have voted for the soup kitchen. Consequently, we can safely make the following assertion about the obligations of a citizen in this situation: if A thinks that the community ought to build a homeless shelter, and thinks that coercive enforcement is necessary and permissible for this end, A ought to vote for the shelter. Assuming that citizens are aware of the obligation to build the shelter and the problems with collective action in the absence of coercion, A will have the beliefs specified in the previous sentence, and thus will be morally obliged to vote for the shelter *regardless* of her beliefs about what other citizens are actually doing or thinking.

In this situation, para-ideal circumstances have been avoided: if each citizen (or enough

---

[28] I thank Lene Bomann-Larsen for raising this point.

citizens to satisfy the voting rule in effect) is properly motivated then, at the conclusion of the vote, the coercive power of the state again endows each citizen with full knowledge of the behavioral plans of other everyone else, bringing it about that each citizen has moral as well as coerced reason to contribute $100. Yet, unlike in the Leviathan case, the advantages of moral expression have been retained. If the vote is conducted in public, each citizen knows that each other citizen correctly perceives the obligation to build the homeless shelter and is genuinely willing to contribute to obeying it (and even to be subjected to coercion for that end).[29] Consequently, EQ-SHELTER once reached, can be expected to persist even if the state ceases to back it up with coercion, since citizens no longer have incomplete information about one anothers' beliefs and intentions. At most, one round of coercion is needed.[30] Also, each citizen's vote at least has the potential to be genuinely morally motivated — a vote for the homeless shelter will ordinarily require and promote honest moral reflection and a commitment to the right. Democracy has allowed us to have the best of both worlds: we have avoided para-ideal circumstances yet permitted collective moral ends to be achieved via the expression of individual moral motivations.

It may be objected that many voting decisions are strategic. The example given is artificial in part because it is a simple one-round pairwise comparison. It is accordingly immune from problems like Arrow's theorem. But we can make slight modifications and again find ourselves in para-ideal circumstances by arranging matters so that voters are faced with a strategic choice. For example, suppose that the voters are called upon to choose between one of three options: a homeless shelter, a soup kitchen, or nothing at all. And suppose each voter realizes that the shelter is the best option, but mistakenly thinks that each other voter supports the homeless shelter with .2 probability, the soup kitchen with .4 and nothing at all with .4. Under such circumstances, each voter has good reason to vote for the soup kitchen in order to make it most likely that the second-best collective moral outcome wins a plurality over the worst outcome. There are many other circumstances, described by social choice theory, under which voting decisions can be strategic and even subject to manipulation.[31]

That problem does not entail that democracy cannot get us out of para-ideal situations, but rather that mere voting cannot work alone. Essential to most contemporary conceptions of democracy are a variety of devices that can resolve strategic problems within voting contexts. For example, there is some empirical evidence that deliberation among citizens can generate single-peaked preferences, allowing deliberating groups to overcome many social choice

---

[29] A necessary condition for this information effect is that the vote provides enough information to reveal that enough citizens to achieve the best equilibrium are committed to that outcome — which can be achieved, by example, by fully public voting, or by a unanimity rule.

[30] In fact, if we are truly in para-ideal circumstances — if every citizen is correctly motivated — then the vote alone should do the trick, even without any coercive enforcement at all, since after the vote each citizen has full information about the motivations of all others.

[31] [citation: Riker]

problems.[32] Likewise, our conception of democracy includes freedom of speech. Ordinarily, we think that free speech serves a variety of traditional functions in a democracy: it ostensibly includes the quality of decision-making, it permits officials to be held accountable, it promotes the discovery of truths, and it is a tool for self-expression, development, and individual and political autonomy. For present purposes, however, it has another function: it permits people to reveal their moral beliefs and behavioral plans, contributing to the elimination of the informational problems that drive cases like the example just given. And likewise again, a democracy with an active civil society associational life permits citizens to evaluate the extent to which their fellows support various positions, and thus avoid the sorts of information shortages that can create para-ideal situations. Much more could be said about this, but the central point should be clear: conventional theories of democracy are not satisfied by voting alone, and neither should a defense of democracy based on its superior strategic properties.[33]

[32] Christian List, Robert C. Luskin, James S. Fishkin and Iain McLean. 2007. "Deliberation, Single-Peakedness, and the Possibility of Meaningful Democracy: Evidence from Deliberative Polls." Working paper, available at http://cdd.stanford.edu/research/papers/2007/meaningful-democracy.pdf .

[33] I have argued elsewhere (in Gowder, "Making Space for Rosa Parks: Democratic Authorship as Political Autonomy," presented at the Public Reason Political Philosophy Podcast Symposium, November 21, 2008, and currently back in working paper form.) that citizen leadership is another essential element of a workable conception of democracy.