RMM Vol. 4, 2013, 71–87
Special Topic: Can the Social Contract Be Signed by an Invisible Hand?
http://www.rmm-journal.de/

Gerald Gaus

Why the Conventionalist Needs the Social Contract (and *Vice Versa*)[*]

Abstract: The recent renaissance of work on conventions, informal institutions, and social norms has reminded us that between the state and individual choice is a network of informal social rules that are the foundation of our cooperative social life. However, even those who appreciate the importance of social norms are reluctant to say that they are about real morality. The first part of the essay examines why this is so. The problem, I suggest, is a widely-embraced view according to which moral judgment is an individual decision about a type of truth that is largely independent of social facts. I show that this popular conception undermines effective social norms and moral conventions. The second part of the essay analyzes the conditions under which effective conventions can be made consistent with diverse individual judgments as to what is morally acceptable—and so conventions can be understood to concern what is genuinely moral. The key, I argue, is the idea of a publicly justified morality as modeled by a hypothetical social contract.

Keywords: social norms, moral conventions, public justification, social contract.

1. Morality and Conventions

I entirely concur with Robert Sugden (2009) on the fundamental importance of evolved conventions in structuring social life. The recent renaissance of work on conventions, informal institutions, and social norms has reminded social and political philosophers that between the state and individual choice is a network of informal social rules that are the foundation of our cooperative social life. Interestingly, however, even those who appreciate the importance of the social norms that we live by are reluctant to say that they are about MORALITY—in capital letters! Norms and conventions are one thing, MORALITY is quite another (e.g., Sugden 1990, 770; Bicchieri 2006, 20–21).

My concern in this essay is, first (*section 2.1*), to analyze why this is so—why even those who truly appreciate the way that norms and conventions are basic to social life, draw back from claims about morality. I shall argue that two, apparently attractive, and certainly widespread, convictions entail that social conventions and norms are distinct from real, true, MORALITY. *Section 2.2* then shows how these two convictions lead to what I shall call the Oracular Conception of

[*] I have greatly benefitted from discussions with Chad Van Schoelandt about these matters, as well as from his comments on an earlier version. My thanks also to Bernd Lahno and an anonymous reader.

Moral Judgment. *Section 2.3* argues that the Oracular Conception undermines a critical function of morality in human social life, and should be rejected. *Section 3* of this essay then considers how a conventionalist understanding of morality can capture the two fundamental insights without lapsing into the Oracular Conception. This, I shall argue, leads us to the Public Conception of morality, as modeled by a hypothetical social contract.

2. The Oracular Conception

2.1 Two Convictions

My recent book *The Order of Public Reason* (2011) has been regularly criticized by orthodox moral philosophers for denying something along the lines of:[1]

> *The Strict Positive / Normative Distinction:* Whether or not Alf morally ought (or ought not) to ϕ (or whether he has a moral right R, or a moral duty D) in society S does not (in any important way) depend on whether there is a convention or an actually recognized social rule in S according to which Alf ought to ϕ (or has a moral right R, or a moral duty D).

As stated, this is rather vague (and I don't think I can do much better). I add 'in any important way' as even the most orthodox of moral philosophers might allow that moral duties sometimes refer to conventions (say, a moral duty to drive on the right in America, but on the left in the UK). But, overwhelming, it is thought that moral imperatives, rights and duties are independent of conventions, as the Strict Positive/Normative Distinction requires. Conventions and social practices, of course, should conform to the demands of MORALITY; failure to do so does not undermine MORALITY, but rather condemns the convention.

The Strict Positive/Normative Distinction is one way to capture a fundamental and necessary feature of morality: it allows us to stand back from current social arrangements and so reflect on their acceptability. The worry about a thoroughly conventionalist view of morality, which ties real morality to positive morality, is that it is apt to be too conservative to be about real morality. Social morality is not only a socially evolved tool for living together, it is a tool that provides us with a critical standpoint to change the terms of our association. The Burkean conservative properly appreciates the fundamental role of evolved traditions in our moral life, but—because he is obsessed by the fragility of these traditions—he is unable to account for the critical stance in social morality. Traditions and conventions are often the result of bias, belief traps, and reputational cascades that can render them highly objectionable (Bicchieri 2006, ch. 5; Sunstein 2006, 122–31). Those who endorse the Strict Positive/Normative Distinction easily make sense of this, for MORALITY is essentially independent of

[1] See Bistagnino, forthcoming.

our socially evolved conventions, and so always can be drawn upon to criticize them.

The second basic conviction is that, in an important sense, conclusions about true morality are matters of individual judgment.

> *The Moral Autonomy Conviction*: Each competent moral agent in society S properly arrives at her own judgment as to what MORALITY requires. The justification of this judgment does not require (*i*) reference to any collective determination, decision, or social fact as to what the morality of society S is, nor (*ii*) does the correct answer to what MORALITY requires depend on what the person inquiring thinks it requires, or what the other people in S think it requires.

The Moral Autonomy Conviction does not require that each autonomous agent make up her mind about MORALITY in isolation; she certainly may consult, and deliberate with, others. But if she does so these are merely inputs into her own decision process, for ultimately she must arrive at her own judgment of what MORALITY requires, and this judgment does not necessarily refer to any collective determination. Clause (*i*) is important since *all* beliefs of a person ultimately depend on her own judgment. Even if social morality were, say, the outcome of a vote, her judgment about it would depend on *her* beliefs about the outcome of the vote. Ultimately, there is no way to escape the point that one's judgments are based on one's beliefs. The Moral Autonomy Conviction insists that not only (of course) is this so, but in coming to this judgment it is not necessary to include any beliefs about what we, collectively, have arrived at as our morality.

Clause (*ii*) is intended to capture the idea that a morally autonomous agent's inquiry into MORALITY aims at coming to a conclusion about a matter the correctness of which is independent of this inquiry and her beliefs about it. She does not think her believing that 'MORALITY requires ϕ', or that the flawless use of her decision procedure has led to the conclusion that 'MORALITY requires ϕ?' *constitutes* the truth of the claim. The Moral Autonomy Conviction insists that MORALITY is an object of inquiry, which is not itself constituted by anyone's inquiry or decision procedure. To render clause (*ii*) more vivid, let us say that each autonomous person believes that there is a MORALITY that is independent of her beliefs about IT, and her inquiry seeks to understand what IT is. We do not together make IT up, nor does she; she inquires into IT, and what IT requires of her and of everyone else. This is not to say that IT must be a real property in the world (though many who have this conviction think IT is), but IT must have a status that is independent of a person's deliberation about IT. What MORALITY requires does not depend on what I think IT requires although, of course, *my conclusions* about what IT requires do most definitely depend on how I see IT. Again, there is no possibility of getting outside of one's beliefs.

Clause (*ii*) also indicates that acting morally is akin to acting on a personal norm (see Bicchieri 2006, 20; Nichols 2010). From my perspective, my obligation to ϕ is independent of what others believe and do; I thus understand my obligation as unconditional on others endorsing the relevant norm (their view of IT). Morality is not fundamentally a collective endeavor. To be sure, it is about what we must all do, but that is determined by IT—as I see IT.[2] Even if everyone lies, I am (and they are) obligated to tell the truth; even if everyone steals, I am (and they are) obligated to respect property.

2.2 The Oracular Conception

Overall, the Strict Positive/Normative Distinction and Moral Autonomy Conviction leads to conception of morality that is highly individualistic in its justification, but collectivistic in its application. Each person properly makes up her own mind about what we all must do. Of course, the person insists that she is not claiming simply that other people should do as *she* instructs, for that would be to claim an illicit authority to dictate others' actions and boss them around. No, it is IT that is requiring us to ϕ: she is simply speaking for IT. Thus the oracular nature of the view; each person takes up the position of an oracle, not only giving her vision of IT, but also telling us that all ought to conform to her vision of IT. For that is what IT—not she!—requires. There is no need to have her judgments confirmed by others; if she concludes that she has deliberated as well as she can, and that she is justified in believing that IT requires ϕ, then that is what she proclaims, and so, in the name of IT, she concludes that all must ϕ.

Hobbes thought he saw the fundamental problem with the Oracular Conception: in the face of disagreement in the use of our reason, for each person to claim the status of the truth giver is to invite civil conflict. To Hobbes, the Oracular Conception exemplifies "intellectual vainglory", an insistence that my opinions about IT are superior to others, so all should be guided by my opinions (Kraynak 1982). The fundamental fact of modern social life, Hobbes observed, is that we disagree about what is "right reason". Each employs his own reason, and concludes that he has reasoned correctly, and so his reason tracks right reason. But this leads to disputes and conflict, which cannot be resolved by anyone's claim to have the correct answer (about what IT is), for our dispute is about who has the correct answer. For Hobbes, someone who insists on the oracular stance shows himself unfit for human society, for he would have us all live by his reason (Hobbes 1994[1688], 23).

2.3 Morality: Descriptive, Social, and Personal Norms

I believe that a version of Hobbes's criticism is sound. Although the Oracular Conception may not lead to overt civil strife, it weakens the normative basis of social cooperation, because it undermines social norms. The crucial distinc-

[2] There are some who hold that claims of morality are only descriptive, and do not entail imperatives addressed to others. The conclusion 'ϕ is unjust' is, they say, simply a moral fact, with no necessary practical implications. I leave aside this thoroughly academic understanding of the moral enterprise though, so long as a person refuses to internalize norms inconsistent with her view of these moral facts, the critical problem raised by these two convictions will apply to this descriptive view.

tion here is between descriptive and social norms. The Oracular Conception of Moral Judgment is entirely consistent with conventions understood as descriptive norms, but it is hostile to conventions that are sustained by social norms. And it is the latter that is fundamental to human social existence.[3]

Consider Robert Cialdini, Carl Kallgren and Raymond Reno's important work distinguishing descriptive and injunctive norms. In their analysis, descriptive norms "characterize the perception of what most people do" while injunctive norms "characterize the perception of what most people approve or disapprove as injunctive norms (or the norms of 'ought')" (1990, 203). Merely descriptive norms often characterize mutually beneficial shared ways of acting, but we do not have normative expectations that people ought or must act in these ways, or hold them responsible for failing to do so. Thus, for example, a signaling system typically functions as a descriptive norm (Bicchieri 2006, 39). In a restaurant it is beneficial to all if customers and wait staff share a convention about what constitutes a signal that one desires the check, but no one is held responsible for employing idiosyncratic, less effective, signals, and generally no one is tempted to do so.

Norms of justice and fairness are not simply descriptive norms characterizing conventions that secure mutual benefit; they are types of social norms, which are characterized by internalization and normative expectations, as well as liability to criticism and other sanctions for violations. Drawing on Cristina Bicchieri (2006, 11; see also Gaus 2011a, 167), let us say that a social norm N exists in group G when a sufficiently large number of members recognize N. More precisely, suppose we have group G divided into subgroups g_1 and g_2. We can say that social norm N exists in G if g_1 is a sufficiently large proportion of G such that for each individual (call such a person Alf) in g_1:

1. Alf recognizes N as norm that applies T circumstances;
2. Alf typically has a motivating reason to conform to N rather than act simply on his own goals in T circumstances on the condition that
   - (a) Alf believes that a sufficiently large subset of G conforms to N in T;
   - ($b1$) Alf believes that a sufficiently large subset of G expects Alf to conform to N in circumstances T or,
   - ($b2$) Alf believes that a sufficiently large subset expects Alf to conform to N in T, prefers that Alf does so, and will sanction Alf for noncompliance.

Such norms function in moralistic ways: we not only expect others to follow them, we think that they 'ought' to; we criticize them for failure to follow them, and are typically indignant or resentful when they fail to do so. More than that, many social norms—in particular *moral norms*[4]—are internalized (Kitcher

[3] In many contexts we would want to distinguish conventions and social norms. For example, a convention may be classified as a type of a descriptive norm, which is to be contrasted with a social norm (Bicchieri 2006, 43ff.). However, in the present context I will use 'convention' to describe structures of interaction that may be based on either. I shall use 'moral convention' and 'moral norm' as synonyms.

[4] These are necessary, not sufficient conditions, for a social norm to be moralistic. For a social norm to be seen as a moral norm, it cannot be seen as highly contingent or arbitrary. Evidence

2011, 93–95; Boehm 2012, 113ff.; Gaus 2011a, 202–205); a person guides herself by them, and feels guilt and shame when she violates them. She expects others to comply and will be resentful and indignant in the face of unjustified violations. To be sure, she often follows them in a flexible sort of way (Boehm 2012, 29ff.). As Bicchieri and Alex Chavez have found, we are apt to be self-serving; in ambiguous situations we focus on the interpretation of the norm that best suits our interests, and will often act against the norm when we can preserve the appearance of complying (Bicchieri and Chavez 2009, forthcoming). Nevertheless, we are genuinely internally motivated, in a flexible manner, to comply with moral norms.

Why all this additional, costly, characteristically moral, apparatus—internalization, guilt, resentment, and punishment? If the aim is simply to coordinate in ways that are mutually beneficial, a descriptive norm will usually do the trick. The heart of social morality is securing cooperation in social dilemmas and mixed motive interactions, where we all do better if we cooperate, but each is tempted to defect while others cooperate (Baier 1995, 292). The ever-present temptation to defect precludes cooperation via simple descriptive norms. Although universal compliance is better for all than universal non-compliance, it is not the case that universal compliance is better for me than my non-compliance when the rest comply. And because of our flexible conscience the equilibrium on rules of social morality cannot even be maintained by interest *and* conscience: a stable social-moral rule requires punishment.

We saw above that the Oracular Conception sees morality as a type personal norm: it is one's personal decision as to what IT is, which is unconditional on the actions of others, though IT (as one sees IT) purports to direct all. Now a society in which most have this, and only this, conception of morality, will find it exceedingly difficult to adopt common social norms, for they require a collective determination of what rules are to be internalized, when I can expect others (or myself) to express guilt and remorse, and when I expect others (or myself) to be sanctioned. As I have stressed, it is not enough to show that such norms are in the interests of all; a person must not only conform to a moral norm, but internalize it, feel guilt, experience the reactive attitudes of resentment and indignation, and be prepared to sanction those who violate it. We now confront a version of Hobbes's point: if each person sees MORALITY through the lens of the Oracular Conception, she will demand that all moral norms correspond to IT as she sees IT. In good conscience, she could not endorse a moral norm that requires her to internalize, feel guilt and remorse, and experience resentment and indignation, unless all these moralistic attitudes are approved by MORALITY as she understand IT. If, given her view of IT, she believes that rule N fails

indicates that even young children distinguish norms that are arbitrary and easily changed (say, by a teacher), such as 'Raise your hand before asking a question' from moral norms such as 'Don't pull the hair of other children', which are not open to revision by authority figures (Gaus 2011a, 122–126). I consider more carefully additional conditions that must be met for a social norm to be a moral norm or rule in Gaus 2011a, 294–303.

to correspond to what IT requires, she cannot internalize N or feel guilty for violating it.

Suppose someone is convinced, say, by John Rawls's argument for the difference principle, and so insists that only property rules that conform to it are consistent with MORALITY, as she understands IT. And suppose she believes that our current property norms are, given the Rawlsian view, unjust. What will be her attitudes towards our current property norms? She might accept that there is no easy way to rid ourselves of them quickly and fear the consequences of a chaotic attempt to do so, and so she might treat current ownership rules as useful descriptive norms, facilitating the coordination of actions. But given her beliefs about their injustice—their failure to correspond to her view of MORALITY—she cannot internalize the rules of property, experience guilt if she disrespects them, and be indignant when she observes violations by others. They are, after all, unjust, and one should not feel guilty about not doing what is unjust.[5]

2.4 Two Views of the Moral Enterprise

A truly conventionalist view of morality understands our moral life as comprised of moral norms or conventions about justice. Morality and justice are *themselves* evolved systems of conventions and norms for living together. For the true moral conventionalist, "orthodox moral philosophy has gotten nowhere because it asks the wrong questions" (Binmore 2005, 1): it inquires about IT, when it should be looking at us, and what we can live with. On this, essentially naturalized view of the moral enterprise, social morality and justice are perhaps *the* innovation that made humans the social creatures we are. (IT did not do it for us, we developed it). We can bring morality down from the heavens, and understand it as the distinctive and perhaps crowning innovation of our species, which enables us to be the types of creatures we are. Like Darwin (2004[1879], 120), the conventionalist is apt to "fully subscribe to the judgment of those writers who maintain that of all the differences between man and the lower animals, the moral sense is by far the most important".

Those enticed by the Oracular Conception resolutely reject this entire picture, and so will not be much worried by my criticisms. They deny that MORALITY has a necessary point or social function, that it is a human innovation that was, and continues, to be a solution to a fundamental problem of social living. To one who adopts the Oracular Conception, MORALITY just *is*, and IT commands us what to do (through her, of course). It is not a tool that we have developed for coordinating with, and controlling others (and oneself). For a surprising number of philosophers (indeed, I fear most), even if no one ever believed, or acted upon, the conviction that we have a duty to ϕ—indeed apparently even if humans were

[5] Of course there may be derivative duties of MORALITY, such as the duty not to unduly upset expectations, which temper this general conclusion. More generally, one can simply report that MORALITY tells us to obey the social norms of one's community; or at least the ones approved by IT. But the problem posed by moral autonomy remains.

a very different sort of species so that *no one would ever ϕ*—it could nonetheless be our unimpaired moral duty to ϕ.[6]

3. The Public Conception

3.1 Reconsidering the Two Convictions

I do not purport to show that this latter picture is definitely false—all these things *could* be true, just as it *could* be true that God could have designed the moral enterprise. As David Gauthier (1991) might say, the universe *could* be enchanted in these ways. But to those of us who are convinced that, first and foremost, morality is an evolved tool for human cooperative life, we must conclude that under the modern condition—where we disagree about right reasoning concerning morality—the Oracular Conception, as a general analysis of moral judgments, undermines the crucial moral norms that underlie our distinctly eusocial existence. Of course on some matters moral judgments may act like the personal norms we have considered (*sections 2.1, 2.3*), but if moral judgments generally satisfy the Strict Positive/Normative Distinction and the Moral Autonomy Conviction, it is very difficult to see how any specific moral norms could stably exist. Assuming that people autonomously arrive at moral judgments, we would expect that many of these personal judgments would not closely align with the norms of the community; consequently, not enough people will endorse the norm as expressing an internal 'ought', and thus the 'enough others' condition (*section 2.3*) for the existence of a social norm is very likely not to be met.[7] Moral reflection thus becomes the enemy of moral conventions, including those concerning justice.

That the Oracular Conception undermines moral conventions is a strong reason to reject it; but the moral conventionalist cannot simply declare victory, claiming that the moral norms of one's community constitute genuine morality. For the crux of the positive/normative distinction does pull us—we *do* think that what is truly moral can depart from the conventions of one's society. And the moral autonomy conviction also pulls us: while it may not be necessary for a moral agent to *always* think things through for herself, there are certainly times when an agent must. A reflective moral agent is one who critically examines her moral conventions in light of her own deliberations about what is valuable, important, and morally acceptable.

We might seem to have hit a dead end: we cannot both respect these two basic convictions and have a realistic hope of maintaining moral conventions. This is too hasty. The problem is not with these general convictions, but the

[6] G. A. Cohen (2008, 20) insists that the infeasibility of a vision of justice does not "defeat the claim of a principle". See also Estlund 2011, defending the relative independence of the demands of justice from our natures.

[7] As we shall see in *section 3.4*, a norm that most see as unjustified can be stable if the participants are caught in 'belief traps'.

rather radical interpretations of them advanced by Strict Positive/Normative Distinction and the Moral Autonomy Conviction. In the remainder of the paper I shall argue that a conventionalist understanding of interpersonal morality (such as the demands of justice), is consistent with two more modest renderings of these core convictions.

While the Strict Positive/Normative Distinction must be set aside, it is critical that we endorse:

> *The Fundamental Normative/Positive Distinction*: That according to the positive conventional morality of society S Alf ought (or ought not) to ϕ (or has a moral right or duty) does not imply that Alf really morally ought (or ought not) to ϕ (or has a *bona fide* moral right or duty) in S.

That social morality is necessarily conventional does not imply that it is merely conventional. The ability to stand back from our conventions, and decide whether they correspond to one's idea of an acceptable way of living together is fundamental to being an autonomous person. Those born into social arrangements, who are indoctrinated such that they cannot question them, but internalize and enforce them just because that is the done thing in their group, are indeed heteronomous: they are ruled by the norms of their group. Thus we should also accept:

> *The Fundamental Moral Autonomy Conviction*: A reflective moral agent in society S properly arrives at her own judgment as to whether the moral norms of her society warrant her internalization, and the adoption of the other moralistic attitudes (guilt, resentment, etc.). (*i*) The moral norms must be congruent with her personal normative convictions, or at least not conflict with—or at a minimum not conflict too sharply with—them. (*ii*) Norms that pass this test are endorsable.

The Fundamental Moral Autonomy Conviction is not so extravagant as our original specification. Autonomy does not require that each come to her own conclusion about what MORALITY requires, independently of the judgment of others.[8] One way we might see a fundamental difference between the two formulations is between optimality and acceptability. The Moral Autonomy Conviction underlying the Oracular Conception identifies moral judgment with a sort of optimality: one believes that 'MORALITY requires ϕ' is justified if it is the best answer to the moral problem, the one that is most coherent, or the one that based on principles that are in in 'reflective equilibrium', for you.[9] Our more modest conviction

[8] Again, this does not mean 'without discussion'; it means that the judgment is in no way a collective one.

[9] The notion of 'reflective equilibrium' is open to multiple interpretations. In Rawls (1999a, 19) the idea seems to be that one's moral convictions fit "together into one coherent view". The suggestion here is that there will usually be a single most coherent view (there will be one unique equilib-

does not suppose that *any* personal judgment that ϕ is the optimal answer justifies belief that morality requires ϕ, for to be required, ϕ must also be part of a socially-maintained convention. But more than that, as clause (*i*) of the Fundamental Moral Autonomy Conviction stresses (which is adapted from Rawls 1996, 11, 40, 140), a morally autonomous person does not require that her society's moral norms perfectly match, or cohere with, her own individual moral reflection about optimality, but only that they are not opposed, or not too deeply opposed, to her normative perspective. Sharing a cooperative social existence under moral norms is a great human good; a morally mature person does not reject the enterprise because it does not perfectly correspond to her own controversial judgment about optimality.

3.2 What is a Genuinely Moral Convention? Public Justification

The moral conventionalist's problem, then, is to show that we can have stable moral conventions that are genuinely moral, and which respect the moral autonomy of all participants. We need to distinguish social norms that have a claim to be 'truly moral' from the wide variety of stable social norms that are oppressive, foolish, and so on, even if our society treats them as moral (Baier 1956, 173ff.). We have seen that the problem cannot be solved by saying, for example, that a truly moral norm is mutually beneficial, for we need to know whether the special features of internalization, guilt, normative expectations and social sanctions for non-compliance are justified. We have to show why a descriptive norm is not enough; if a descriptive norm will achieve mutual benefit, relying on the costly machinery of a social norm looks like a type of collective folly.

Recall that for the conventionalist view of morality to succeed, the solution to our problem cannot itself imply that equilibrium on a norm we all recognize as a *bona fide* moral norm is well-nigh impossible (*section 2.3*). If social-moral norms are to perform their function of ordering social life in acceptable ways, there must be a reasonable prospect that a group can achieve an equilibrium on a moral norm that all see as normatively acceptable given the exercise of their moral autonomy, and that such an equilibrium could be a matter of public knowledge. This is basically what Rawls (1996, 387) has called "public justification": the public knowledge that a stable practice is morally acceptable to all the participants.

The idea of public justification opens up the possibility of reconciling our two fundamental convictions (*section 3.1*) with stable moral convention. Consider:

> *The Public Justification Principle*: If a (positive) moral convention C in society S is endorsable by all participants, C is a genuinely moral convention.

rium, so we do not face a multiple equilibria problem). I explore and criticize this view in Gaus 1996, 101–108. However in Rawls 1999b (289), he rejects the idea that reflective equilibrium is necessarily a matter of coherence. "Reflective equilibrium requires only that the agent makes these revisions with conviction and confidence, and continues to affirm these principles when it comes to accepting their consequences in practice."

Notice that the Public Justification Principle employs the idea of the endorsable (clause (ii) of the *Fundamental Moral Autonomy Conviction*) as the basis of determining whether a convention is genuinely moral; if all participants, reflecting on what they consider morally acceptable, have sufficient reason to endorse C, then C is not only a positive convention, but one that *really* is moral. Now note that if this test is satisfied, the convention will tends towards stability insofar as participants' reflection on C tend to induce affirmation of it; each rational participant realizes that she has sufficient reason to internalize it. Publicly-justified moral norms incline towards stability, as they will be widely internalized among the reflective or those acting in accordance with their reasons. Rather than undermining the social function, moral reflection now reinforces it. Not a trivial result.

3.3 The Diverse Social Contract: Modeling Public Justification

The social contract is best understood as way to model public justification. It seeks to gain insight into our justificatory problem by modeling it as a deliberative one (Gaus 2011a, 264ff.; Rawls 1999a, 16). The Public Justification Principle is an evaluative principle; the social contract provides a procedure that seeks to satisfy it. Let me briefly sketch a version of this procedure that I have developed elsewhere (Gaus 2011a, ch. 5).

Since our fundamental problem is moral and valuational disagreement, we must commence with a group of individuals who not only know they disagree about these matters, but who know what they think about them. If, as did Rawls, we model them in a way that brackets their disagreements so that they only know that about which they agree, there will be severe worries about whether the agreement will be stable once we look at real parties, who know their full set of evaluative standards. Unless the relevant convention is endorsable given a person's total set of evaluative considerations, it would not be rational for her to internalize it and adopt the relevant moral attitudes. As is well-known, this problem plagued the Rawlsian project right to the end.

So let us assume a group of contractors who are members of group G, or we might say, Members of the Public, $P_1 \ldots P_n$, who are fully aware of their evaluative convictions. How to fix the membership in set? Recall that support by a sufficiently large group, g_1 (a subset of G) is needed to maintain a social norm (*section 2.3*). Thus a minimally effective social contract must at least include a group large enough to sustain the norm. If at least g_1 does not see the convention as justified, it is apt to fail (without extensive coercion). Supposing that g_1 is a proper subset of G, the remainder group, g_2, is not, we assume, necessary to maintain the convention. The reasons to expand the set of 'Members of the Public' well into g_2—ideally to all of G—go beyond basic sustainability to reasons concerning the efficiency of equilibria, and the ability to maintain them without extensive coercion. A contract that omits from $P_1 \ldots P_n$ a subset of G (say g_2), it likely to yield a contractually justified norm that is not endorsable by g_2. Although g_2 may not be absolutely required to sustain the convention, they would nevertheless be subject to it, at least in the sense that they would be coerced for noncompliance. Their inability to acknowledge the demands of the convention as legitimate in the face of temptations to defect, will impair social relations between them and their fellows: "Only if such acknowledgement is possible can there be a true community between persons in their common practices; otherwise their relations will appear to them as founded to some extent on force." (Rawls 1999b, 59)[10] And indeed significant force will be required, for g_2 will not internalize the convention. The convention will require more force to sustain itself, and defections will be more common. Moral exclusion thus has significant costs. For these and other reasons the liberal ideal of the social contract has sought to extend the normative justification of current moral practices as widely as possible—ideally, to all normal competent members of the moral community.

Assume, then, that $P_1 \ldots P_n$ approaches G. We can model each member of $P_1 \ldots P_n$ as initially exercising something akin to the Oracular Conception's view of moral autonomy, proposing the convention that she deems optimal. This would generate a set of proposed conventions, which each member of $P_1 \ldots P_n$ then orders.[11] Given moral diversity, it is implausible to suppose that their orderings are identical, or even that they agree on an optimal element. If they did, the Oracular Conception would, after all, be consistent with effective social norms, for we happen to converge on optimality.[12] Given that we are almost certain to discover significant disagreement, including at the top of each person's ordering, each party must then resort to our more modest conception of moral autonomy, and consider which proposals are endorsable by her. The set of proposals that are endorsable by everyone in $P_1 \ldots P_n$—let us call this the *socially eligible set*—represents the set of publicly justified conventions, all of which can see as morally acceptable, and which will be stable because everyone in $P_1 \ldots P_n$, which we suppose approaches G, can internalize them. No one would have deep moral objections to any of these conventions. They would not be condemned by anyone in $P_1 \ldots P_n$ as oppressive or immoral.

The idea of the social contract thus gets real moral leverage in a morally diverse community that cannot agree about the best arrangement, but can secure an 'overlapping consensus' on some member of a family (set) of alternative arrangements (Rawls 1996, 133–173). The point is most definitely *not* that an arrangement is normatively justified because the parties agree to it. Just the opposite: given the way we model the agents, their 'agreement' to a convention is the way we understand the claim that, from the perspective of each, it is normatively endorsable. The point of contractual argument is to show that a moral practice could be stable 'for the right reasons'—the social network that maintains the moral norm could do so in light of their full set of normative com-

[10] In Gaus 2013 I argue that the reactive attitudes will be undermined, but this is a more complex matter.

[11] More formally, I show that an incomplete quasi-transitive ordering is sufficient (Gaus 2011a, ch. 5).

[12] I try to show in Gaus 2014 that such convergence is not even an ideal, in the sense of a good thing for a society.

mitments. If we can assure ourselves of that, we would have confidence that the moral conventions that structure our social cooperation are not simply the dead weight of past, and present, oppression, but are endorsable by all, given what each cares about.

3.4 Why Actual Consent Won't Do

Actual consent or actual agreement plays no role in this account of the social contract (though, of course, actual choices of populations play a pivotal role in the evolution of a convention, see *section 3.5*). The point of the social contract is to get moral distance from current conventions and social norms, which, *pace* many admirers of moral conventions, are not maintained simply by free action, exchange and common interest, but also by internalization, guilt and punishment. We need to ask ourselves whether our conventions are simply the products of force, ignorance, taboo, or inertia, or whether they can be normatively justified to all.

Now many libertarian philosophers, including Anthony de Jasay (2010), insist that only actual consent, or real agreement, could legitimize submission to a coercive institution—and, as I been stressed, a moral norm is such an institution. Indeed, following Rawls we might use a politically-flavored term, 'public moral constitution', to describe the basic social framework of moral relations in a community (Gaus 2013). For the quintessential consent theorist a person is bound to comply with a public moral constitution (or moral convention) if and only if she has (genuinely) agreed, promised, or somehow voluntarily submitted. The literature had rehearsed many times the difficulties of this approach; let me simply stress two that are less often mentioned which are especially germane to our topic.

(*i*) Because, as we saw in *section 2.3*, Alf's compliance to moral norms heavily depends on his beliefs about the expectations of others (and of course each person in the norm's relevant network occupies such a position), it is entirely possible, and all too often is the case, that norms that in fact people have universally agreed to are ones that many, perhaps all, find normatively unacceptable. Particularly important are the ideas of pluralistic ignorance and belief traps (Bicchieri 2006, ch. 5). Consider a real-world social norm, such a female genital cutting, which has been carefully studied by Bicchieri (2014). One might assume that in areas where the practice predominates it is approved of by most—especially males—and is generally enforced on females. But as Bicchieri reports, in some countries with very high participation rates in the practice, surveys report very high disapproval rates. These data are consistent if, as seems to be the case, in these areas most people have the belief that others expect them to engage in the practice, and that the others' normative expectations will be disappointed if they do not. Fathers may normatively disapprove of cutting their daughters, but believe that other fathers will not allow their sons to marry uncut females. And this might well be true, for even fathers of sons who disapprove of the practice could believe that their neighbors will sanction men who marry uncut females. Thus the vast majority might consent to cutting, sanction those who do not participate and so on, while the vast majority also deeply normatively disapprove of the practice. When consent to a norm depends on complex beliefs about the normative expectations of others, and what they believe that you believe about them and so on, social norms that—from the perspective of the members of the group—are normatively non-endorsable may obtain universal consent. The social contract steps back from the fact of agreement, asking whether the distinctive normative standards of the parties gives each reason to prefer participation in this practice to its abolition, which is very different from whether, given the actions of others, one will consent to go along.

(ii) I have been stressing that many moral norms are critical to solving social dilemmas; we have good reason to cooperate, but when others cooperate each has a reason to defect. The textbook case is, of course, the Prisoner's Dilemma. Now one of the things that Hobbes taught us is that any attempt to promise our way out of Prisoner's Dilemmas lands us in another Prisoner's Dilemma, in which the dominant strategy is to make the promise and break it. Of course Jasay (2009, 402) recognizes the problem and points to a familiar solution: "In some of these game-like interactions, a free rider option is a temptation to deviate from the equilibrium. Since ascertainable historical and empirical facts testify to the survival of such conventions, there must be enough 'players' using contingent retaliatory strategies to spoil the free rider payoff and keep deviation in check."

There are two ways to understand the important proviso that "there must be enough 'players' using contingent retaliatory strategies to spoil the free rider payoff and keep deviation in check." On the first interpretation, which is predominant in the rational choice literature, Betty is committed to retaliating because it is part of her equilibrium strategy in an iterated mixed-motive game with Alf. The foundation of this interpretation is the folk theorem, which shows that in a variety of mixed-motive iterated games the players have cooperative equilibrium strategies so long as each can punish the other for defection, driving the other down to the payoff he/she would get if the punisher played the minimax strategy (Binmore 2005, 79ff.). Recent modeling has shown, however, that the information demands of such reciprocal dyadic strategies are enormous once the group size becomes even moderately large (say beyond 30–50) (Gaus 2011a, 87ff.; Bowles and Gintis 2011, 63–76). Each needs a bookkeeping system keeping track of the distinct history of individual transactions with each other member of the group. Very little modeling of the evolution of cooperation in large groups, where information is included as a scarce resource and individuals make mistakes about their partner's history, employs such dyadic retaliatory strategies.

The alternative interpretation, which is at the heart of more recent modeling and is supported by extensive experimental data, is that the stability of social-moral norms depends on a widespread readiness of participants to punish any violations they observe. These are altruistic punishers: Betty acts to punish any observed violator, regardless of whether she is apt to receive individual returns from the punishing act (Bowles and Gintis 2011; Gaus 2011b). This is indeed fundamental to norm stability, but it has little to do with the view of norms as a type of self-interested exchange. When Betty punishes she does not consult the violator's history to see if he has promised to obey the norm; she is concerned with what he has just done. And she is explicitly *not* concerned with best promoting her own interests (she is an *altruistic* punisher). She observes a violation of a moral norm, and the normative reactions of resentment and indignation are evoked; because she is normatively invested in the norm, she is ready to expend resources upholding it. We are thus led away from interest, exchange and consent to the normative basis of the convention—and that is precisely what social contract theory focuses on.

3.5 The Limits of Justification and Coordinating on a Convention

I have argued that a social contract justification makes sense when there is a non-singleton set of morally eligible conventions or norms. However, so powerful is the theoretical—and perhaps the moral—urge for determinacy that social contract theorists, even when recognizing that individuals have reasonable disagreements as to what is the best arrangement, typically employ some procedure that selects one option as the socially best from the eligible set. Among many others, Gauthier (1986, ch. 5) and Binmore (2005, ch. 2) employ bargaining theory to transform individual utility conflicts into a uniquely correct way to decide how much utility each person is to receive from the contract; in our terms, we must understand this as a proposed unique solution as to how much a moral convention should satisfy the normative commitments of each participant. As Sugden (1990) has demonstrated, given the dependency of bargains on expectations, there is no determinate outcome to such bargaining problems. More generally, one thing that the last forty years of inquiry has taught us is that all these procedures are highly controversial. The irony of the typical social contract theorist is that, basing her entire edifice on the fundamental insight of normative disagreement, she then claims that we would agree on the best way to resolve it. As Nozick (1974, 98) observed in another context, "[w]hen sincere and good persons differ, we are prone to think they must accept some procedure to decide their differences, some procedure they both agree to be reliable or fair. Here we see that the possibility that this disagreement may extend all the way up the ladder of procedures."

A social contract theorist who does not give into the urge to resolve moral disagreement by trotting out a controversial resolution procedure recognizes that the contract device is properly indeterminate: it yields a set of socially eligible arrangements, but cannot tell us which to choose. And so again we must appeal to Hume and the evolution of conventions. We can understand each moral norm in the socially eligible set as a possible equilibrium in an impure coordination game. We disagree about what is the optimal coordination point (moral convention), but because each sees every member of the socially eligible set as a convention which he or she has good reasons to normatively endorse, each rationally prefers coordination on any one to the failure to coordinate on what all see as a possible *bona fide* moral convention. For achieving some eligible moral convention is a great good; as Rawls would say, we achieve a stable cooperative social life for the right reasons. Which eligible convention we settle on will be largely a matter of history and accident. Sugden (1990, 769–770) saw this long ago: "Contractarianism [. . .] cannot generate a uniquely impartial code of morality. If contractarian reasoning can generate any moral conclusion at all, it must be that morality is, in important respects, a matter of convention." But, as I have argued, we should not follow Sugden's other suggestion, that "contractarian reasoning leads not to morality but to norms" (1990, 770).

The social contract philosopher is apt to recoil at this—within the bounds of the justifiable, what social morality we end up with is a matter of contingency and happenstance, not the unique dictates of an impartial procedure. But that is the lesson that the Humean has always been trying to teach us, if only we listened.

References

Baier, K. (1958), *The Moral Point of View: A Rational Basis of Ethics*, Ithaca: Cornell University Press.

— (1995), *The Rational and the Moral Order: The Social Roots of Reason and Morality*, La Salle: Open Court.

Bicchieri, C. (2006), *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge: Cambridge University Press.

— (2014), *Norms in the Wild*, Cambridge: Cambridge University Press.

— and A. Chavez (2009), "Behaving as Expected: Public Information and Fairness Norms", *Journal of Behavioral Decision Making* 22, 191–208.

— and — (forthcoming), "Norm Manipulation, Norm Evasion: Experimental Evidence", *Economics and Philosophy*.

Bistagnino, G. (forthcoming), "Gerald Gaus and the Task of Political Philosophy", *European Journal of Analytic Philosophy*.

Binmore, K. (2005), *Natural Justice*, Oxford: Oxford University Press.

Boehm, C. (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.

Bowles, S. and H. Gintis (2011), *The Cooperative Species: Human Reciprocity and Its Evolution*, Princeton: Princeton University Press.

Cialdini, R. B., C. A. Kallgren and R. R. Reno (1990), "A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior", *Advances in Experimental Social Psychology* 25, 201–34.

Cohen, G. A. (2008), *Rescuing Justice and Equality*, Cambridge/MA: Harvard University Press.

Darwin, C. (2004[1879]), *The Descent of Man*, second edition, New York: Penguin.

Jasay, A. de (2010), "Ordered Anarchy and Contractarianism", *Philosophy* 85, 399–403.

Estlund, D. (2011), "Human Nature and the Limits (If Any) of Political Philosophy", *Philosophy & Public Affairs* 39, 207–235.

Gaus, G. (1996), *Justificatory Liberalism: An Essay on Epistemology and Political Theory*, New York: Oxford University Press.

— (2011a), *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, Cambridge: Cambridge University Press.

— (2011b), "Retributive Justice and Social Cooperation", in: White, M. D. (ed.), *Retributivism: Essays on Theory and Practice*, Oxford: Oxford University Press, 73–90.

— (2013), "On the Appropriate Mode of Justifying a Public Moral Constitution", *The Harvard Review of Philosophy* 19, forthcoming.

— (2014) "The Role of Conservatism in Securing and Maintaining Just Moral Constitutions: Toward a Theory of Complex Normative Systems", in: Levinson, S. and M. Williams (eds.), NOMOS: *Conservatism*, New York: New York University Press.

Gauthier, D. (1986), *Morals By Agreement*, Oxford: Oxford University Press.

— (1991), "Why Contractarianism?", in: Vallentyne, P. (ed.), *Contractarianism and Rational Choice*, Cambridge: Cambridge University Press, 15–30.

Hobbes, T. (1994[1688]), *Leviathan*, with selected variants from the Latin edition of 1688, ed. Edwin Curley, Indianapolis: Hackett.

Kitcher, P. (2011), *The Ethical Project*, Cambridge/MA: Harvard University Press.

Kraynak, R. P. (1982), "Hobbes's Behemoth and the Argument for Absolutism", *The American Political Science Review* 76, 837–847.

Nichols, S. (2010), "Emotions, Norms, and the Genealogy of Fairness", *Politics, Philosophy & Economics* 9, 275–96.

Nozick, R. (1974), *Anarchy, State and Utopia*, New York: Basic Books.

Rawls, J. (1996), *Political Liberalism*, paperback edition, New York: Columbia University Press.

— (1999a), *A Theory of Justice*, revised edition, Cambridge/MA: Harvard University Press.

— (1999b), "Justice as Fairness", in: Freeman, S. (ed.), *John Rawls: Collected Papers*, Cambridge/MA: Harvard University Press, 47–72.

— (1999c), "The Independence of Moral Theory", in: Freeman, S. (ed.), *John Rawls: Collected Papers*, Cambridge/MA: Harvard University Press, 286–302.

Sugden, R. (1990), "Contractarianism and Norms", *Ethics* 100, 768–86.

— (2009), "Can a Humean Be a Contractarian?", *RMM* 0, *Perspectives in Moral Science*, ed. by M. Baurmann & B. Lahno, 11–23.

Sunstein, C. (2006), *Infotopia: How Man Minds Produce Information*, Oxford: Oxford University Press.