Self-organizing Moral Systems:
Beyond Social Contract Theory[1]

"But what if morality is created in day-to-day social interaction,
not at some abstract mental level?"
~Frans de Waal

Gerald Gaus

1 TWO MODES OF MORAL THINKING

A common view of moral thinking — perhaps most characteristic of moral philosophy — understands reasoning about moral claims to be, in a fundamental sense, akin to reasoning about ordinary factual claims. On this commonsense approach, when Alf deliberates about a moral claim or demand (say, that people ought to ϕ), Alf considers the best reasons as he understands them for and against the claim, including what he takes to be the correct normative principles, perhaps checks his conclusions with others to see if he has made any errors, and then comes to the conclusion, "we all ought to ϕ." His moral reasoning may refer to facts about other people (say, their welfare), but it is not a general requirement on the moral reasoning of any competent agent that he always takes as one of his reasons the moral deliberations of others. To be a little more precise, we can identify:

> *The "I conclude we ought" View:* As a competent moral agent, if (*i*) Alf conscientiously deliberates and concludes that, given what he takes to be the correct normative premises and relevant empirical information, one ought to ϕ (ought not ϕ, or may ϕ) under conditions *C*, where this does not require taking account of the conclusions of the deliberations of others and (*ii*) he reasonably concludes that morality instructs that we all ought to ϕ (ought not ϕ, or may ϕ) under conditions C, then (*iii*) he ought to ϕ in circumstances *C*.

It is important that on the "I conclude we ought" View Alf does not typically assert that we all ought to ϕ in *C because he* has concluded that we ought to ϕ: Alf may believe that "we ought to ϕ" in *C* because it is a moral truth that we ought to ϕ, or that an impartial spectator would approve of our ϕing. The important point is that once Alf conscientiously comes to the belief that one ought to ϕ in *C* — it is, we

[1] I am especially grateful for detailed comments by Jack Knight and Jon Riley on an earlier version of this paper.

might say, his best judgment about the morally best thing to do — then, as a competent moral agent, he will justifiably ϕ in circumstances *C*, and indeed insist that we all do so, for that is what we ought to do.[2] And, as I have stressed, none of this necessitates (though it may be epistemically recommended) that Alf factors into his moral deliberation the conclusions of others.

On one reading the social contract offers another view of justice and morality. Hobbes, Locke, Rousseau and Kant all hold that individuals' "private judgments" about morality or justice radically diverge, and because of this individual private judgment is an inappropriate ground — or at least, I shall argue, an inappropriate *sole* ground — for demands of justice. Kant famously insists that, even if we imagine individuals "to be ever so good natured and righteous," when each does what "seems just and good to him, *entirely independently of the opinion of others*" they live without justice.[3] This apparently paradoxical conclusion — that a world of people who acted only on their own sincere convictions about justice would live without justice[4] — derives from two commitments of social contract theory. (*i*) It is taken as given that reasoned private judgments of justice inevitably conflict. This is partially because of self-bias, but only partially: innate differences in emotional natures, differences in beliefs which form the basis of current deliberations, differences in education, socialization and religious belief — all lead to pervasive disagreement. (*ii*) Secondly, it is assumed that a critical role of justice in our social lives is to adjudicate disputes about our claims and so coordinate normative and empirical expectations. For Kant the problem of universal private judgment was that "when there is a controversy concerning rights (*jus controversum*), no competent judge can be found."[5] Each, thrown back on her own reasoning, ends up in conflict, and ultimately unjust relations, with others. Understood thus, a necessary role of justice (or morality) is to provide an interpersonally endorsed adjudication of conflicting claims.[6] Securing justice, on this second view, is

[2] I assume here that conditions *C* are so defined that typical justifications for not ϕing (duress, etc.) would show that *C* was not met. For present purposes we can set aside these complications.

[3] Kant, *The Metaphysical Elements of Justice*, 2nd edition, edited and translated by John Ladd (Indianapolis: Hackett, 1999), p. 116 [§43]. Emphasis added.

[4] I have defended this paradox in some depth in "The Commonwealth of Bees: On the Impossibility of Justice-through-Ethos," *Social Philosophy & Policy*, forthcoming.

[5] Kant, *The Metaphysical Elements of Justice*, p. 116 [§43]. Emphasis added.

[6] See John Rawls, "An Outline of a Decision Procedure for Ethics" in *John Rawls: Collected Papers*, edited by Samuel Freeman (Cambridge, MA: Harvard University Press, 1999): 1-19.

inherently something we do together.[7] If no other good-willed and conscientious moral agent accepts that in circumstances *C* justice demands ϕ, Alf's demand will not secure just social relations.

Given points (*i*) and (*ii*), social contract theorists have endorsed a *Joint Reasons Requirement*: in some way a *bona fide* claim of justice in society *S* must be based on, take account of, or be a function of, the different reasons of members of *S*, or their different deliberations about justice. The paradox of private judgment about justice, the social contract theorist insists, can only be overcome by an appeal to the Joint Reasons Requirement. In section 2 I further analyze the Joint Reasons Requirement, distinguishing two versions, the Collective and the Social. I shall argue that traditional social contract theory stresses the Collective Interpretation. However, the Collective Interpretation ultimately endeavors to supplant "I conclude we ought" reasoning with the Joint Reasons Requirement, but to so thoroughly set aside one's own ("I conclude we ought") moral reasoning undermines one's status as a free moral agent. I argue that the Social Interpretation is superior: it accommodates the basic thought behind the Joint Reasons Requirement while acknowledging the importance to each agent of his own moral deliberations. With the Social Interpretation defended and explicated, section 3 then takes some initial steps in understanding how the Social Interpretation replaces the traditional collectivistic model of the social contract by examining several simple models of self-organizing moral systems of free agents. Section 4 reflects on the nature of choice-based models of morality, trying to clarify their aims and limits. Finally, section 5 offers a few comments on the potentially radical implications for moral thinking of refocusing the Joint Reasons Requirement toward self-organizing moral systems in diverse societies.

2 SOCIAL THINKING ABOUT JUSTICE

2.1 Justice as a Social Property

The worry about relying exclusively on "I believe we ought" reasoning is that my conclusion is about the justice of a joint action — what we do — but as a competent moral agent my deliberations control only what I do, not what others do, so I alone cannot produce the joint action. Consequently my "I believe we ought" judgment is

[7] It is not only contract theorists who think this. See G. A. Cohen, *Rescuing Justice and Equality* (Cambridge, MA: Harvard University Press, 2008), pp. 175ff.; R. B. Brandt, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979), chap. 9.

very often ineffective in securing what I believe we both ought to do.[8] Alf has concluded that, say, "we ought to both perform action α," but the question remains whether he ought to α if Betty refuses to α. In this case, whether or not Alf's "I believe we ought" judgment is action guiding even for Alf depends on how he ranks the alternative joint actions in terms of justice. Contrast, for example, the interactions modeled in Figures 1 and 2.

		Betty α	**Betty** β
Alf	α	4th / 1st	2nd / 2nd
	β	3rd / 3rd	1st / 4th

FIGURE 1: "I BELIEVE WE OUGHT ϕ" IMPLIES "I OUGHT TO ϕ"

		Betty α	**Betty** β
Alf	α	2rd / 1st	3rd / 3rd
	β	4th / 4th	1st / 2rd

FIGURE 2: "I BELIEVE WE OUGHT ϕ" DOES NOT IMPLY "I OUGHT TO ϕ"

In the interaction of Figure 1, each orders the outcomes: (1) we both do what (on my view) is just; (2) I do what is just and the other acts on the inferior view; (3) I act on the inferior view and the other acts on the better view (at least someone does!); and (4) we both act on the inferior view. In this game the sole equilibrium is that Alf acts on his view (α), and Betty acts on her view (β), of justice. At either of the coordination solutions (when both play α or both play β), one of the parties would do better by changing his or her move, and acting on his or her favored

[8] If, as do some, we suppose that judgments of justice are not intended to be action guiding, this is not a problem. I consider the extent to which judgments of justice are inherently practical in *The Tyranny of the Ideal: Justice in a Diverse Society* (Princeton: Princeton University Press, 2016), pp. 11-18. Here is it simply assumed that our concern is action-guiding judgments; if there are other notions of justice, they raise different issues.

interpretation of justice. So here even if the other does not do as you have concluded "we" ought, you still ought to do it.

Figure 2 does not support this conclusion. There are two equilibria in this classic impure coordination game: both act on α and both act on β. Here both parties agree that from the perspective of securing just social relations, it is best that they do not necessarily act on their understanding of optimal justice, for if they do so they may secure an inferior justice. Now insofar as the aim of justice is securing social relations of a certain moral quality, we would expect Figure 2 would be the typical interaction. When "I believe we ought to α" implies "I ought to α," my action alone is sufficient to infuse adequate justice into our relations, though it would be more just if you too did as I concluded you ought. But there are many interactions where your action can undo the effect of mine in securing just social relations. Consider Figure 3:

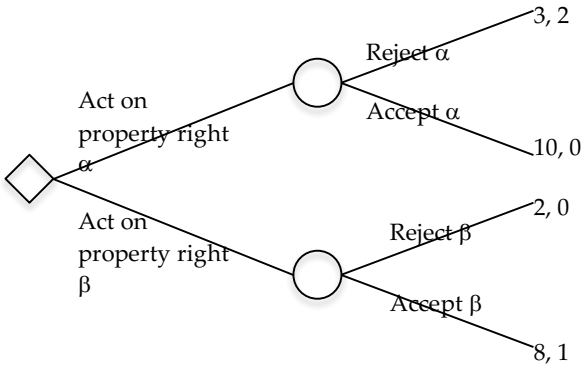

FIGURE 3: A SEQUENTIAL PROPERTY RIGHTS GAME

Alf moves first (at the diamond), then Betty responds (at the circle choice nodes). On Alf's view property rights scheme α is the most just; if, as he believes, they both ought to act on it, he would score it 10 out of 10 on justice. The β property rule is, he judges, less just; if they both acted on it he would score it 8 out of 10. Betty however, believes not only that the β right is more just, but that α is unjust, and she prefers to reject it (utility = 2) than to acquiesce (utility = 0). If, however, Alf must act on a property right that Betty rejects, it would be better to act on the more just α (3) than the less just β (2). In this game the sole subgame perfect equilibrium is that Alf and Betty both act on β. From the perspective of each, justice demands that one take account of the action of the other, and so Alf's "I believe we ought to α" judgment does not entail an "I ought to α" judgment.

I suppose here that judgments of justice are typically about interactions along the lines of Figures 2 and 3, not 1. Judgments of justice are about securing a certain type of social relation and, especially among large groups of people, the individual's unilateral action seldom can secure this quality. It is this sense in which, I suppose, justice is a social-moral good. As Plato stressed in the *Republic*, it is about relations among individuals rather than unilateral action; that is why the theory of justice has been a part of social and political philosophy right from the beginning. Much moral thinking can be described as simply "I believe *I* ought" reasoning, where my only concern is what I ought to do, come what may. On can be chaste, honorable and honest alone; one cannot make promises, keep contracts, or determine mutual expectations about what is proper and improper on one's own. Here the moral theorist switches from "I believe I ought" to "I believe we ought" judgments, but once we have made that move, there is still a problem: I have concluded what we ought to do, but I cannot secure this without your cooperation. That is why theories of justice — even Rawls's — can helpfully employ game theory, which is a theory of strategic interaction, modeling what we both will be led to do.[9]

2.2 Individual and Collective Reasoning in Contract Theories

We thus come to appreciate that determining the rules and institutions of justice is — at least to some extent — a social problem. A natural interpretation of this, which we might see as definitive of the social contract tradition, is to understand it as a collective problem. My concern here is not to present a thorough criticism of this response — which, I think, has offered, and continues to offer, fundamental insights[10] — but to clearly contrast it to the alternative I shall develop.

Consider first a *substantive* version of the social contract, presenting a theory that suitably-characterized individuals would agree to common principles of justice to structure their social and political lives. Rawls's contractualist theory is, of course the quintessential case. Such substantive theories seek a "we believe we ought to" view. The theory constructs an account of what reasonable and rational persons with certain motivations *would* agree what we should all do. The theorist,

[9] See John Rawls, *A Theory of Justice*, rev. edn. (Cambridge, MA: Harvard University Press, 1999), pp. 237-8, 303-7, 505. For the importance of strategic considerations in Rawls's theory see Paul Weithman, *Why Political Liberalism? On John Rawls's Political Turn* (New York: Oxford University Press, 2010).

[10] For important recent contributions, see Michael Moehler, *Minimal Morality: A Two-level Contractarian Theory* (New York: Oxford University Press, forthcoming); Peter Vanderschraaf, *Strategic Justice* (New York: Oxford University Press, forthcoming).

then, actually presents something like an "I believe that [we believe we all ought to]" view of justice. That is, the theorist provides an analysis of what *she* thinks *we* would collectively agree to as common, shared, principles of justice to regulate *our* relations. Rawls suggests an even more complicated view. As he presents his theory of justice, "you and I" develop a theory of what all reasonable people would agree to.[11] We thus seem to have something like an "I [Rawls] believe that [<you and I believe that> we believe we all ought to]" view.[12]

Now a theory that acknowledges that we disagree about justice, yet need to coordinate, is certainly a great improvement upon simple "I believe we ought to" reasoning, as it seeks to confront at a basic level the fundamental moral insight that unless you and I concur about the demands of justice, our social relations will be deeply flawed from the perspective of justice (as in Figure 2). Yet, at the end of the day, it is a theory of what the theorist (Rawls) believes that you and he believe that we all believe what we all ought to do. That is, at the end of the day, it is one person's conviction about what we all believe we ought to do; and for the same reasons we disagree in our simple "I believe we ought" judgments, we disagree in our "I believe we believe we ought" judgments.[13] Rawls, indeed, came to recognize that reasonable people do disagree about the most reasonable conception of justice — about the conception of justice that we would agree to.[14] Rawls's own theory of "justice as fairness" — a theory about what we will agree on — is highly controversial. And even if two philosophers, Alf and Betty, accept Rawls's principles of justice, they are almost certain to disagree on their interpretation, leading them to interactions along the lines of Figure 3.

At the meta-level of what we all agree to, substantive social contract theories thus also struggle with the problem of disagreement of private judgments about justice.[15] In response, many traditional social contract theories adopted a *procedural* solution: if Alf and Betty find themselves in a reasonable dispute about the demands of justice, they should appeal to some umpiring procedure that generates

[11] John Rawls, *Political Liberalism*, expanded edn. (New York Columbia University Press, 2005), pp. 28ff. See also Rawls, *A Theory of Justice,* p. 44.

[12] It is for this reason that Habermas is correct to characterize Rawls's theory as "monological" at the highest level. Jürgen Habermas, "Reconciliation Through the Public Use of Reason: Remarks on John Rawls's Political Liberalism," *The Journal of Philosophy*, vol. 92 (March, 1995): 109-31 at p. 117.

[13] Or "I believe that [<you and I believe that> we believe we all ought to]" judgments.

[14] Rawls, *Political Liberalism*, p. xlvii.

[15] See further my "Public Reason Liberalism" in *The Cambridge Companion to Liberalism*, edited by Steven Wall (Cambridge: Cambridge University Press, 2015): 112-40.

a common, salient, answer, on which they can coordinate.[16] Such proceduralism elevates formal institutions with adjudication mechanisms to the role of arbiters of justice: if the institutional procedure selects α as the demand of justice, then the procedural social contract view is that public reason instructs that "α is demanded by justice." Rawls insists that, at least in its most expansive form, this proposal cannot be accepted: "The conception of political justice can no more be voted on than can the axioms, principles, and rules of inference of mathematics or logic."[17] Even if we are reluctant to go that far, and so we accept that sometimes a vote determines justice in our polities, surely it remains correct to insist that this public, procedural, reason cannot displace one's own "I believe we ought" reasoning about justice.[18] Even if the majority dictates that "we believe we ought to α," a free moral agent, exercising her capabilities, may nevertheless conclude that "*I* believe we ought *not* α," and so reject "public justice." Recall in Figure 3, where Betty concludes that it is better not to coordinate at all than to coordinate on property right α. At least some forms of democratic proceduralism — for example, Rousseau's — are powerful explications of "we believe we ought" reasoning, but tend to accord too little weight to a person's "I believe we ought" judgments. To be sure, each citizen has a vote to express her "I believe we ought" judgment, yet when it comes to justice, to oneself one's own conclusions almost always count for more than 1/*N*, where *N* is a very large number. A plausible account of justice as a social property must give due regard both to a person's "I believe we ought" reasoning and to coordinating with the justice judgments of others. With matters of justice Luther was right: sometimes one must declare that one's conscience prohibits accommodation.

Another version of the social contract, combining substantive and procedural elements, takes this problem seriously, seeking a reasonable balance between the two modes of reasoning. Abstracting from the many fascinating technical details, the heart of this approach is to identify two points for each individual, say Alf and Betty: the justice-based "payoff" that one would receive from unilateral action based on one's "I believe we ought" reasoning (the so-called "no agreement point") and the best "payoff" one could receive from coordinated action. We suppose both

[16] I explore this approach in *Justificatory Liberalism: An Essay on Epistemology and Political Theory* (New Your: Oxford University Press, 1996), Parts II and III.

[17] Rawls, *Political Liberalism*, p. 388n.

[18] Cf. David Gauthier, "Public Reason," *Social Philosophy & Policy*, vol. 12 (Winter 1995): 19–42. I explore this problem in "The Property Equilibrium in Our Liberal Social Order (Or How to Correct Our Moral Vision)." *Social Philosophy & Policy*, vol. 28 (Summer 2011): 74-101.

would gain "moral utility" (better moral outcomes)[19] through some form of coordination; if this was not the case, one would not agree to coordinate, and we would have the interaction modeled in Figure 1. In the toy example in Figure 2, the no-agreement point (α, β) is each person's third option; it is the best outcome that each could receive if, as it were, they walked away from an agreement to coordinate. Both would gain by moving to (α, α) or (β, β). The question is which is to be chosen. Suppose we employ utility measures where 10 indicates an option that perfectly satisfies one's "I believe we ought" standards of justice, while 0 represents an option which the agent sees as unacceptable from the standpoint of justice. Assume that the relevant utilities in a cardinal version of Figure 2 (the impure coordination game) are (always reporting Alf's first): (α, β) = (2, 3); (β, β) = (10, 7); (α, α) = (10, 5). For each, either coordination outcome is thus superior to the best non-coordination case, but they differently evaluate the justice of the two coordination outcomes. Alf judges (α, α) to be perfect from the perspective of justice, while Betty only judges it to be 5 out of 10. On the other hand, Betty judges (β, β) to be perfect, and Alf scores it 7 out of 10. Drawing on bargaining theories that focus on division over a distributable good such as, say, money or time, some social contract theorists suggest we might see Alf and Betty's problem as deciding how much utility it would be rational for each to give up to secure a bargain. On the most common approach, axioms are defended that identify a unique division of the utility gains.[20]

Formalizing decision problems in terms of utility representations can make many issues clearer, as I hope to show in the next section. As Rawls recognized, even deontological theories such W.D. Ross's can be faithfully represented in terms of standard cardinal utility measures.[21] But this important insight by no means licenses forgetting about what the numbers are representing, and so treating all

[19] It is critical to keep in mind that "utility" is simply a mathematical representation of a person's ordering of states of affairs, not itself a good which is sought. To say that an agent maximizes utility is simply to see her as one whose actions are directed to obtaining the state of affairs that she ranks as best. If her rankings are based solely on moral criteria, then, assuming that common formal consistency conditions are met (e.g., if α is better than β, then β is not better than α) are met, her choices can be represented in terms of maximizing (moral) utility: she has a moral utility function.

[20] See, for example, John Nash, "The Bargaining Problem." *Econometrica*, vol. 18 (1950): 155-62; Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986), chap. 5; Ehud Kalai and Meir Smorodinsky, "Other Solutions to Nash's Bargaining Problem," *Econometrica*, vol. 43 (1975): 513-18; Ehud Kalai, "Proportional Solutions to Bargaining Situations," *Econometrica*, vol. 45 (1977): 1623-30. For a noncooperative formulation, see Ariel Rubinstein, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, vol. 50 (1982): 97-109.

[21] Rawls, *Political Liberalism*, p. 332n.

"utility" as essentially a homogenous abstract quality subject to the same types of disputes and resolutions. If we do not keep in mind that the utility scores represent disagreements about justice, it may seem that Alf and Betty have a resource division or interest compromise problem, with the numbers indicating unequal splits of the "gains." Clearly this is not the case. Neither party has approved of the other's understanding of justice, and may in fact have severe doubts about it. Suppose at (α, α) Betty appeals to Gauthier's principle of minimax relative concession and complains that at (α, α) she must make a 5/7 relative concession while as (β, β) Alf would only make a 3/8 relative concession.[22] Fair's fair, after all; and the maximum concession should be minimized.[23] Alf replies: "A basic reason why you concede relatively more at (α, α) than I do at (β, β) is that you place too much value on living according to your preferred rule of justice and too little on sharing; I "concede relatively less" because I place so much more value on sharing rules than do you, even those rules I consider to be an inferior, while you more highly evaluate living according to your favored rule. You are free to do so, but I fail to see why that decision gives you a claim to additional consideration, such that I should move toward your preferred rule just because you (in my view) erroneously undervalue sharing a rule with others."

Alf disagrees with Betty about justice, though given his view he also holds that justice is enhanced when he and she can share a rule, so he is prepared to move away from his optimum rule if that will increase sharing an approximation of justice. Note that all this is about justice *from his own perspective*, which includes a commitment to his own insights about perfect justice and the importance of living according to shared rules. *These are already factored into his utility function.* There is

[22] According to minimax relative concession, we compute relative concession according to the formula:

$$\frac{u(fp) - u(fc)}{u(fp) - u(ip)}$$

where *u(fp)* is one's "first proposal" (the most which one could claim from the bargain without driving the other party away (for both of them this is 10); *u(fc)* is what one actually receives from a bargain [at (α, α) this is (10, 5), at (β, β) this is (7, 10)] and so represents one's "final concession"; I suppose that *u(ip)* is the utility of one's initial position — the utility one comes into the agreement assured of, in this case (2, 3).

[23] Some has argued that common bargaining axioms are only about rationality, and have no implications concerning fairness. I believe this is wrong; motivations for the critical symmetry axiom invoke a general notion of equality. See John Thrasher, "Uniqueness and Symmetry in Bargaining Theories of Justice," *Philosophical Studies*, vol. 167 (2014): 683-699. Interestingly, in his final statement of his view Gauthier sees minimax relative concession not as a solution specifying a rational bargain, but as a standard of justice. See his "Twenty-Five On," *Ethics*, vol. 123 (July 2013): 601-24.

no more value of sharing to be taken into account — all the information has already been captured in his utility function. Consequently, that his own conception of justice leads him to take account of what rules he can share with others in no way commits him to treating his and other people's views of justice as somehow "on par" and each having a symmetric claim to the "moral gains" produced by shared rules. On reflection, I think that assumption looks rather bizarre.

3 SELF-ORGANIZATION IN MORALITY

3.1 From Constructivism to Spontaneous Orders

It may help to pause and take stock of the analysis thus far. I am assuming here that the focus is on informal moral rules or rules of justice, say R_1 and R_2. The fundamental claim has been the importance of two modes of reasoning about justice (and, more broadly, social morality). When employing "I believe we ought" reasoning a moral agent inquires, given her standards of justice, as to the extent to which R_1 and R_2 express the way we ought to act in our social interactions. Given her understanding of justice, I suppose that she can score every rule from 0 to 10, where 10 indicates perfect agreement with her standards and 0 indicates that she judges that the rule is simply unacceptable as rule of justice. A critical assumption here is that a moral agent acknowledges acceptable approximations to her understanding of perfect justice (scores of 1-9).[24] The reason that she would accept living according to an imperfect rule is supplied by recognition of the social character of justice: justice cannot be fully secured by unilateral action. Just social relations require participation of others, and so one's complete understanding of justice will ultimately weigh the relative importance of "I believe we ought" considerations and joint reasoning, which indicates rules that we can share.

We have seen that the social contract tradition takes seriously the social nature of justice, providing accounts of how we might collectively reason about shared rules to live by. In an important sense, however, these accounts are — with perhaps a few exceptions[25] — what Hayek would call forms of "constructivism."[26] They remain versions of "I believe that [<the set of reasonable deliberators believe that>

[24] See further my *The Tyranny of the Ideal*, pp. 45-9 and "The Commonwealth of Bees." We could formalize these into von Neumann – Morgenstern utility functions, but nothing critical turns on this point.

[25] Ryan Muldoon's work is, I think, the most obvious exception to this generalization. See Ryan Muldoon, *Social Contract Theory for a Diverse World* (New York: Routledge, 2016).

[26] F. A. Hayek, *Rules and Order* (London: Routledge, 1973), chap. 2. Cf. Rawls, Political Liberalism, Lecture III.

we all ought to]" judgments, in which the theorist constructs her version of what joint reasons endorse, or the reasonable compromise between our moral views. It is not, I think, going too far to say that they are "top-down" (from the theorist to us) theories of what a "bottom-up" (what we collectively would choose) morality might look like. Once we recognize this, the question looms: what would a genuine "bottom-up" morality look like? Such a morality, I shall argue, would be the result of each person acting on what might be called, more than a little inelegantly, her "I believe what, taking account of the justice-based choices of others, we ought to do — and that's what I ought to do" view of justice. Here there is no collective choice: the theorist seeks to inquire what would occur if each agent genuinely followed her own view of justice, taking into account her commitment to coordinating.[27] Would agents who disagree in their convictions about (*i*) optimal justice and (*ii*) the relative importance of coordinating on justice, converge on common rules, or would they each go their own way? Under what conditions might free individual moral reasoning replace the collectivist constructivism of the social contract? These are the questions I shall now begin, in an admittedly simplified and tentative way, to pursue.

3.2 Justice-based Utility Functions

The following analysis is purely formal, and does not presuppose any specific substantive theory of justice or social morality. Each person is assumed to be solely interested in acting justly, as she sees it (see §4 for worries). Each is thus characterized by a justice-based utility function that represents her judgments as to the justice of states of affairs, defined in terms of what rules of justice are acted upon. Recall that — at least as I have characterized it — the aim of social contract theory is to see how free and equal moral persons who do not agree about optimal justice can share a common system of moral rules or principles. Substantive-contract attempts to secure this, we have seen (§2.2), valorize a specific conception of justice that the theorist claims "we all [could] share," but this claim itself succumbs to the problem of reasonable pluralism. Some good-willed moral agents who wish to share a system of justice with others do not accept the theorist's conclusions about what "we believe we ought to do." Consequently, I seek here a theory of shared moral life that does not presuppose the correct answer, but

[27] I have considered in some depth the relation of this theoretical perspective to the choices of individual moral agents in "Social Morality and the Primacy of Individual Perspectives," *The Austrian Review of Economics*, DOI 10.1007/s11138-016-0358-8.

instead endeavors to model how individuals who disagree about the correct answer might nevertheless converge on a common moral life.

Given the analysis thus far, I assume that the justice-based utility of each person can be divided into two parts. What I shall call person *A*'s (aka Alf's) *inherent* (justice) utility of rule R_1 (denoted $\mu_A(R_1)$) represents Alf's scoring of R_1 in terms of how well it satisfies his "I believe we ought" reasoning about justice. The following analyses suppose that this ranges from 0 to 10 for all agents, with 0 representing a rule that the agent judges to be unacceptable from the perspective of justice. These utilities do not support interpersonal comparisons. Now agents who recognize the social dimension of justice also place weight on whether others share a rule. A rule R_1 where $\mu_A(R_1) = 10$, but is shared by no one, will be seen by Alf as inferior to R_2, where $\mu_A(R_2) = 9$ but is shared by all. However, again reasonable disagreement asserts itself. Good-willed competent moral agents reasonably differ on the relative importance of the two modes of moral reasoning. Some place greater weight on their "I believe we ought" reasoning (a rule's inherent utility) while others put more emphasis on the social dimension of justice. This is a critical difference that must be at the core of our analyses. To take account of moral disagreement (i.e., about inherent justice-based utility) while imposing a uniform weighting of these two modes of reasoning would miss a fundamental source of our moral differences.[28] We thus suppose that each person has a weighting function that takes account of how many others act on a rule — how widely it is shared — and the importance to her of that degree of sharing. This weighting function for person *B* (aka Betty) will be denoted as $w_B n(R_1)$, which is the weight that Betty gives to Rule R_1 when n others act on it. We suppose weights vary between 0 and 1.

There are an infinite number of weighting systems. Figure 4 presents the ones on which we shall focus, here with $n = 101$.

[28] In this sense the present analysis presses beyond that in *The Order of Public Reason* (Cambridge: Cambridge University Press, 2011).

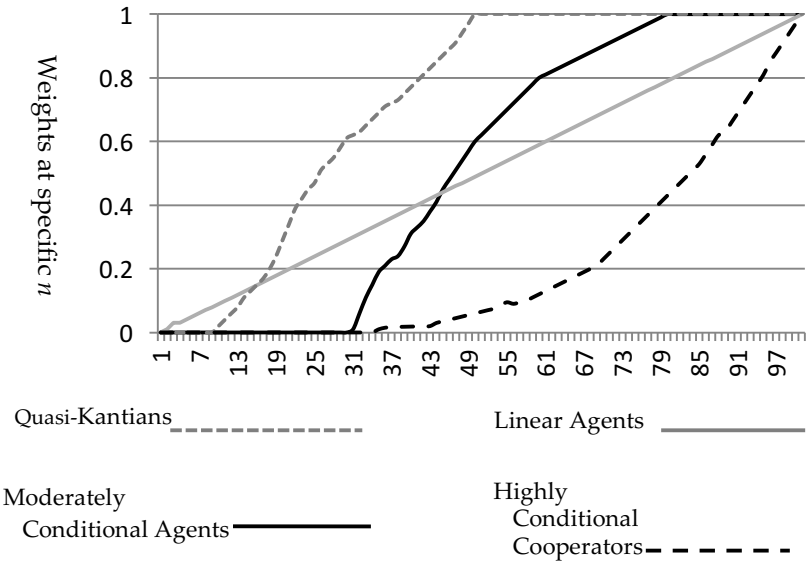

FIGURE 4: FOUR WEIGHTING SYSTEMS

All agents put some emphasis on sharing, thus they admit some pull of the social dimension of justice. "Quasi-Kantian" agents recognize some importance to sharing; they give no weight to a rule that is not practiced by 10% of the population; by 50% they give a rule a maximal weighting of 1. We might say that the act on the categorical imperative – *if* some others do too! Moderately Conditional Agents have similar shape to their weighting function as Quasi-Kantians, but they are more typical of Humean conditional cooperators: until a significant share of the population acts on a rule they are not willing to act, and so give it a 0 weight.[29] A rule must have 30% uptake before they give it any weight, and reaches a maximal weight at 80%. Both Quasi-Kantians and Moderately Conditional Agents, we might say, seek a moral community but not a maximally large one. Linear Agents have, unsurprisingly enough, a linear weighting function: the more share the merrier, but as long as someone else acts on their rule, they give it some positive weight. Lastly, Highly Conditional Cooperators are resolute in stressing the importance of sharing, and only weight rules highly when the large

[29] Cristina Bicchieri models conditional cooperators as having a certain threshold, often requiring that "most" others share a rule before they will act on it. Our types do not have abrupt thresholds, but Moderately Conditional Cooperators have a significant threshold of about 30%. See her *The Grammar of Society* (Cambridge: Cambridge University Press, 2006), pp. 11ff.

majority has already joined in. These Highly Conditional Cooperators are, as it were, willing to play the justice game only if most others do. Highly Conditional Cooperators give no weight to a rule unless about a third of the population follows it, and give very little weight to any rule practiced by less than 60%. They do not give really high weights until approximately 90% practice it. They are thus *highly* conditional moral agents. Highly Conditional Cooperators are roughly the mirror image of our Quasi-Kantians, perhaps died-in-the-wool Humeans.

In our analysis, then, an agent is concerned with both his own evaluations of the inherent justice of a rule (i.e., his "I believe we ought" conclusions) and sharing just social relations with others ("we believe we ought"), and will ultimately make his decision based on his own view of the inherent justice of the rule given his evaluative standards and the weighted number of others who are acting on the rule. Letting U_A be Alf's total justice-based utility of acting on rule R_i, $\mu_A(R_i)$ the inherent justice-based utility of R_i given Alf's evaluative standards, w_A his social weighting and n the number of people acting on R_i, we get:

$$\text{EQ. 1} \qquad U_A(R_i) = \mu_A(R_i) \times w_A n(R_i)$$

If Alf is confronted by two rules, he will act on that which maximizes U_A. So Alf acts on R_1 rather than R_2 only if $U_A(R_1) \geq U_A(R_2)$.[30]

3.2 Agent-based Modeling and Parametric Rationality

Gauthier usefully distinguishes two types of rational choice contexts. In "*parametric* choice…the actor takes his behavior to be the sole variable in a fixed environment. In parametric choice the actor regards himself as the sole center of action. Interaction action involves *strategic* choice, in which the actor takes his behavior to be but one variable among others, so that his choices must be responsive to his expectations of others' choices, while their choices are similarly responsive to their expectations."[31] As with most theoretical distinctions, this one is perhaps not quite so crystal clear as it first seems, but it highlights an important difference in rational choice. In the toy games we considered in section 2.1 Alf's action was dependent on what he thought Betty was going to do and what she was going to do depended on what she thought he was going to do. Indeed, ultimately Alf's choices depended on considerations such as what he thought she thought he thought she was going to do, and so on. Thus the crucial importance of common knowledge assumptions in game theory, whereas in situations of parametric choice a person takes the actions

[30] And he *will* act on R_1 if $U_A(R_1) > U_A(R_2)$.

[31] Gauthier, *Morals by Agreement*, p. 21. Emphasis in original.

of others as a given: what is her best course of action given the actions of others? In contrast to strategic situations she supposes that her choice will not affect the choices of others; their choices are taken as a parameter — a given or constraint — in her decision. Thus, as Gauthier recognizes, rational choice of a consumer in a large market is quintessentially parametric — she simply adjusts her action to the given prices; compare a traditional shopping bazaar in which she is a price-taker *and* setter, and so is engaging is strategic interaction. It is important that in parametric contexts an individual still may seek an equilibrium result: i.e., one in which given the context in which she finds herself she cannot unilaterally increase her utility by a change in her choice.

As I said, as with most distinctions this one becomes less clear as we inspect it more closely. In a large market at any given time each consumer is a price taker, not a price setter, but over a series of periods in the market each consumer's parametric choices affect the price, and are minute influences in setting the price in subsequent periods. So it is an exaggeration to say that one's choices have no effect on others' choices.[32] The critical point is that at each choice node, one takes the actions of others as an exogenous variable that is simply a parameter in one's maximizing decision. It is also something of an exaggeration to say that in parametric choice expectations about others do not matter, as such expectations may be a critical parameter. If one is on one's way to the airport and wishes to minimize travel time by choosing the fastest route, one confronts a parametric choice, but of course one's empirical beliefs about which road is presently congested (one's empirical expectations about how much traffic one will encounter on each route) is critical. Parametric choice requires empirical beliefs about what others are doing, so that one can efficiently respond.

Moral and political philosophers tend to underappreciate the insights of parametric analysis. Indeed, even Gauthier, who has a much deeper appreciation of it than most, holds that morality only enters when strategic choice arises.[33] Strategic choices are often small-number interactions, and so "ppe-inclined" philosophers, often wedded to the strategic outlook, have repeatedly analyzed morality in terms of small-number interactions. We know, though, that a social system of justice is typically a large-numbers phenomenon, in which many

[32] This is a familiar point in agent-based models of adaptive system: "…as agents adjust to their experiences by revising their strategies, they are constantly changing the context in which other agents are trying to adapt." Robert Axelrod and Michael D. Cohen, *Harnessing Complexity* (New York: Basic Books, 2000), p. 8.

[33] Gauthier, *Morals by Agreement*, p. 21. But cf. pp. 170-1.

individuals interact, each adjusting her action to what she sees as the current social parameters. As generations proceed, the aggregation of parametric choices changes the social parameters (as with prices in large markets), which in turn changes what is parametrically rational. Here parametric models can be most enlightening, which suppose that each person has a utility function (a representation of a preference ordering)[34] and beliefs about the state of the world at time t (which typically include what others are now doing), and each acts to maximize her utility. In some agent-based parametric models the system will stabilize in the sense that, at some iteration i, further states of the system confront all the members with exactly the same parameters such that no one henceforth changes their choices (leaving aside preference change, errors in beliefs, and factors exogenous to the system). In other systems there can be endless adjustments; they will never reach system-wide equilibrium.

All the following analyses are agent-based models of this sort. Each agent has a very simple binary option set of acting on either R_1 or R_2, which are assumed to be alternative rules of justice over some area of social life, say property rules, promising rules, privacy rules, etc.[35] It is also assumed that whether a person has acted on R_1 or R_2 in the last period is reliable public information; in period i, each has a correct first-order belief about the actions of others in period $i - 1$. Our aim is to get some preliminary insights as to when free moral agents, each with her own distinctive justice-based utility function and fully committed to acting on it, are apt to converge on a shared rule of justice, and under what conditions they are less likely to.

4 SOME DYNAMICS OF SELF-ORGANIZATION

4.1 The Basic Convergence Dynamic

The important feature of the following models (and this is basic to most agent-based models) is that because of their weightings, individuals parametrically choose in one period; the aggregation of parametric choices in period i can change the parameters in period $i + 1$, leading some individuals to change their choices in $i + 1$. Again, changing parametric choices (as in evolving markets) should not lead us

[34] Again, it is important to always keep in mind that a preference is simply a binary relation according to which one state of affairs (action, etc.) is better than another; it is not a reason or explanation as to *why* one is better than another. To say that Alf holds that "α is preferred to β" because that "is his preference" is a tautology, not an explanation. A utility function is a mathematical representation of a consistent preference structure.

[35] There are some difficulties in further formalizing this idea. See *The Order of Public Reason.* pp. 267ff.

to confuse these interactions with the strategic, interdependent, choices of classical game theory.[36] To better understand the core dynamic that can lead free moral agents to converge, consider a very simple case. Suppose we have a group *G* of agents with all four types of weighting functions equally represented, but only one rule R_1 is deemed eligible [$\mu(R_1) > 0$] by the entire group *G;* another rule R_2 is deemed not only eligible, but inherently superior, by a significantly smaller subgroup, *g* [i.e., for *g*: $\mu(R_2) > \mu(R_1) > 0$; for *G*-*g*: $[\mu(R_1) > 0 = \mu(R_2)]$. If (*i*) interactions within *G* are uniform so that each interacts with each, (*ii*) there is reliable knowledge of the previous actions of *G*, and (*iii*) the distribution of social weighting functions, *w,* among *g* is varied, the entire group *G* has a very strong tendency to converge on the eligible rule, R_1. A bandwagon effect is apt to occur because of diverse weighting systems expressing the value of sharing, and R_1 can be much more widely shared (by all of *G,* not just *g*). To see this consider *A*-types in *g*, say Moderately Conditional Agents, who place great weight on sharing with over 50% of the population, and who see R_2 as only slightly better than R_1. Given the greater number acting on R_1, *A*-types are very likely to decide that they best promote their justice-based utility by switching to R_1. Thus in the next period, the number of those acting on R_2 will no longer be *g*, but *g* minus *A*-types. Now consider *B*-types. Perhaps they were simply Linear Agents who scored R_2 higher than R_1, and in the previous period acted on R_2. However, given the defection of *A*-types, *B*-types now see that they will best promote their justice-based utility by also switching to R_1. And so on, as each type whose total justice-based utility marginally favored R_2 in period *i* comes to reevaluate the parameters it faces, and switches in $i + 1$ to the new maximizing choice, R_1. Eventually we will come to *Z*-types, perhaps Quasi-Kantians. These agents tend to act on the rule favored by their inherent utility (μ) so long as a modest number of others act on it, but at some point even they will give very little weight to R_2, and so defect to R_1.

It is important that this dynamic does not require an entirely smooth distribution of $\mu(R_1) - \mu(R_2)$ differences. What is required is that at each period enough people recalibrate which rule is best from their normative point of view such that, at the next stage, more of those who were still optimizing by R_2 adjust their actions to R_1, until all do. What the dynamic certainly does depend on, though, is that most of the *g* R_2 advocates significantly weight the importance of the number of others with whom they interact on shared rules. If a large

[36] Evolutionary game theory is a more complicated case; replicator dynamics can make it look as if players are strategically responding to each other. They model, I think, essentially evolutionary adaptive systems, which are my concern here.

proportion of g essentially only cares about their inherent evaluations of R_2, they obviously will not adjust their moral behavior.

4.2 The Fully Random Model

Such is the basic logic. We need to get some idea, however, how sensitive the convergence dynamic is to distributions of differences in inherent utilities and weightings. We start with a population of 101 agents, with $\mu(R_1)$ and $\mu(R_2)$ scores between 1 and 10 randomly assigned. Thus all agents hold each rule is an acceptable approximation of justice, if only barely (a score of 1). Agents were randomly assigned to one of our four weighting types.[37] In the first period each individual simply acts on her "I believe we ought" judgment, maximizing her evaluative utility (μ). In ties [i.e., $U(R_1) = U(R_2)$]) an individual acts on R_1; perhaps R_1 is the simpler rule, and so individuals choose it in a tie. Our empirical updating rule is simple, if somewhat dumb: at each period an agent calculates whether in the previous period she would have achieved more utility if she had acted on the alternative rule; if she would have she switches in this period.

As Figure 5 shows, Rule 2, randomly favored by 51 agents compared to R_1's 50, went to fixation after five periods. It might be wondered whether any specific weighting type was crucial in producing convergence, but as Figure 6 shows, under this same population, any three of the weighting systems (again, randomly assigned) resulted in fixation on R_2, giving some reason to believe that the convergence dynamic is not highly sensitive to specific types. However, we do see that combinations of types certainly have an effect; omitting the Highly Conditional Cooperators slowed down convergence.

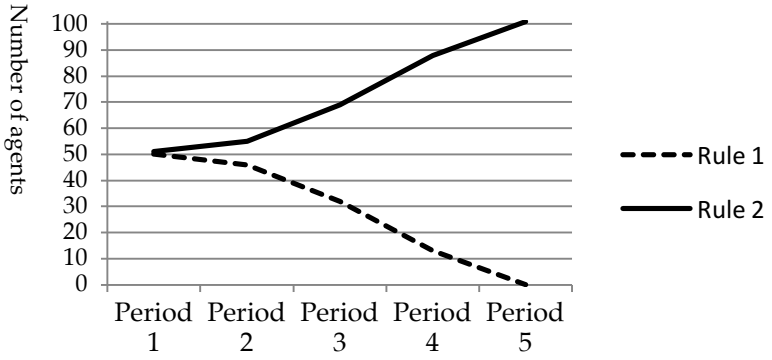

FIGURE 5: CONVERGENCE IN A FULLY RANDOM POPULATION, FOUR WEIGHTING TYPES

[37] In Figure 4. For precise weightings, see the Appendix,

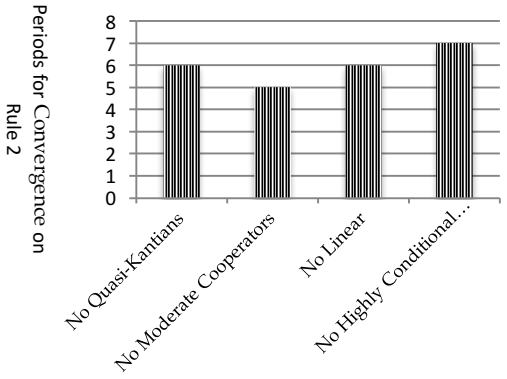

FIGURE 6: PERIODS FOR CONVERGENCE IN FULLY RANDOM POPULATION EMPLOYING COMBINATIONS OF THREE WEIGHTING TYPES

4.3 Moderate Polarity Models

We commenced with a fully random model to explore the core dynamics under conditions where the population was very closely divided, and to better see some of the effects of the different weighting systems. It certainly is clear that the dynamic does not depend on one rule having an initial overwhelming advantage. Different weightings have different thresholds and values, which help induce convergence dynamics. However, fully random distributions of inherent utility and weighting functions are hospitable to cascades, since they tend to ensure that there will be continuity of degrees of $U(R_1) - U(R_2)$ differences, such that whenever one agent switches, this will decisively affect the choice of the next agent "in line," who then switches in the next period, and so on. The question is the extent to which a convergence dynamic applies in non-random populations. An especially difficult case is a polarized population, divided into two mutually exclusive groups, one subgroup thinking highly of one option and scoring the other low, with the other subgroup doing the opposite.

To explore this possibility our group of 101 agents was divided into two "High-Low" groups, one of which scored R_1 between 10 and 6, and R_2 between 4 and 1 [thus $\mu(R_1)$ are all "high," while $\mu(R_2)$ are all "low"]; the other group was assigned the opposite "High-Low" inherent utilities. Each polarized group had approximately an equal division of all four types of agents. The suspicion that polarization makes convergence more difficult was confirmed; very closely split [52 agents $\mu(R_1)$ High; 49 agents $\mu(R_2)$ High] polarized populations did not display tendency to converge. Somewhat surprisingly, perhaps, convergence on R_1 did occur within four periods at the close but not finely balanced [56 agents $\mu(R_1)$ High;

45 agents $\mu(R_2)$ High]. Again, as Figure 7 shows, it was found at this 56/45 division convergence occurred with any three types and again the omission of Highly Conditional slowed down the process (taking 10 periods). Highly Conditional Cooperators appear to excel at jumping into a cascade and completing it. Quasi-Kantians, on the other hand, tend to reinforce the split; they have their maximal weightings at around 50% of the population, and so tend to reinforce polarity. Quasi-Kantians who find themselves in subgroups who agree with them do not easily switch rules. Perhaps the truly striking thing is that even they can be induced to leave their High-Low groups and converge on a common rule, and do so when any two of the three other weighting systems are well represented. If High-Low polarity is not too finely balanced, then, it can be overcome; the diversity of weighting types speeds up the process, inducing sufficient continuity in the populations' $U(R_1) - U(R_2)$ differences even with the population is characterized by polarized (thus discontinuous) $\mu(R_1) - \mu(R_2)$ differences.

It is, I think, worthy of emphasis that as we add diversity of weighting types, we can overcome polarization in inherent justice judgments. Those of us who have been deeply skeptical of claims that, once we filter out biases "we share a common conception of justice," must acknowledge that not only do we disagree about justice, but for the last hundred and fifty years western societies have been significantly polarized between "right" and "left" justice, with the last forty adding a number of other groups (e.g., feminists, environmentalists) who also tend to "High-Low" judgments. Rather than reasonable pluralism we should, perhaps, be thinking of moderate reasonable polarity. Our Polarity Model gives us some reason to suppose that these sharp inherent justice differences can be significantly moderated by a diversity of weighting types. An uptake of this idea would constitute a fundamental change in the orientation of the public reason project, which has thus far supposed that diversity is the problem, and commonality the route to sharing. Here, we see the possibility that one type of diversity can counteract the centrifugal tendencies of another. This is a critical insight. So far from heterogeneity always be an impediment to convergence on a shared rule of justice, some configurations of diversity can help secure agreement. The issue is not "do we agree enough to live together?" but "does the overall pattern of homogeneity and heterogeneity induce convergence on common ways of living together?'

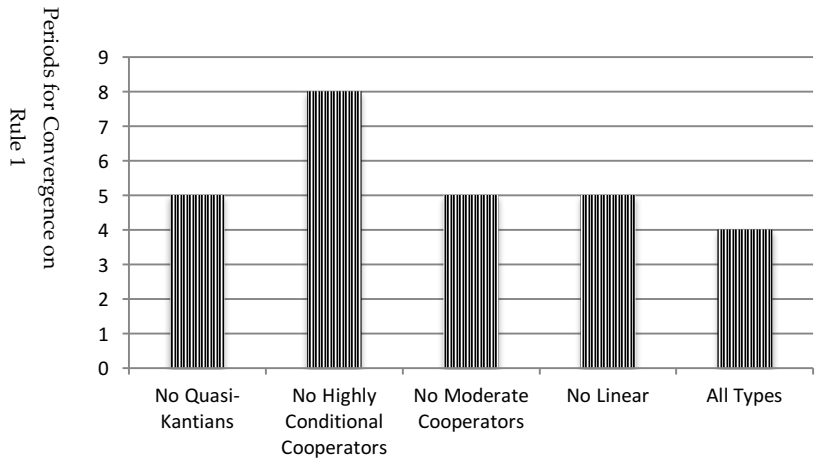

FIGURE 7: PERIODS FOR CONVERGENCE IN HIGH-LOW POLARIZED GROUP

4.4 Differential Reference Group Models

A simplification in the models thus far discussed is that each person takes the entire group as her reference group: at each period she adjusts her action to what the entire group is doing. But often people are concerned with narrower reference groups.[38] Alf might concern himself with what those in his traditional cultural group are doing while Betty's concern is with the actions of those in her urban and work environments. In previous work I have supposed that we seek a practice of moral accountability based on shared moral rules with the widest feasible set of other moral agents.[39] But in many contexts people might be committed to a practice of accountability only with those with whom they regularly interact, while others may be interested in a practice of accountability based on shared rules with some other group. In this case the different elements of the population would have different reference groups — different groups of people with whom they value sharing a moral rule. Can there be shared rules by moderately polarized groups under such circumstances?

To take some first steps in understanding the effects of different reference groups on convergence, let us analyze a somewhat challenging case: the population is not only split into different reference groups, but some of the reference groups display opposite High-Low polarity. In one reference group 3/5 of the population has High-Low evaluative utilities in favor of R_1 (the other 2/5 of the group has

[38] Cristina Bicchieri extensively examines the role of reference groups in her *Norms in The Wild* (Oxford: Oxford University Press, 2016). I made some preliminary remarks on this point in *The Tyranny of the Ideal*, pp. 184-87.

[39] See *The Order of Public Reason*, pp. 279-83.

High-Low evaluative utilities in favor of R_2), while the other reference group has just the opposite High-Low division. Here, we might think, convergence within each group will occur, but not between them: our polarized population models in section 4.3 indicate that with such splits we should expect convergence on the most popular rule in each group. And of course that is what would normally happen if these are entirely unrelated reference groups, for then we simply have two independent populations. The interesting case concerns populations with overlapping reference groups. In the Differential Reference Group Model a population of 150 agents is divided into three main groups, with two of the groups having subgroups. They are:

Group A (50 agents): split population, not High-Low (an agent may have any combination of $\mu(R_1)$ and $\mu(R_2)$ between 1 and 10).

Group B1 (25 agents): approx. 3/5 High-Low favoring R_1; 2/5 High-Low favoring R_2.

Group B2 (25 agents): approx. 3/5 High-Low favoring R_1; 2/5 High-Low favoring R_2.

Group C1 (25 agents): approx. 3/5 High-Low favoring R_2; 2/5 High-Low favoring R_1.

Group C2 (25 agents): approx. 3/5 High-Low favoring R_2; 2/5 High-Low favoring R_1.

Note that both subgroups in B have identical evaluative utility distributions, as do both subgroups in C. The difference is their reference groups, as indicated by Figure 8:

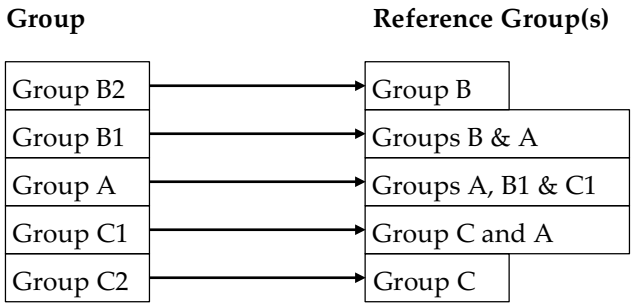

FIGURE 8: DIFFERENTIAL REFERENCE GROUPS

Group B2, then, updates only in relation to the choices of Group B in the previous period; this means that Group B2 (*i*) has a reference group of 50 agents and (*ii*) their entire reference group has a 3/5 High-Low bias toward R_1. B1's reference group is 100 agents, encompassing all of Groups A and B. Because Group A is composed of our random inherent utility agents, Group B1's reference group includes both B2, which shares its High-Low bias toward R_1, and the random group. Group A, the random group, has a reference group of 100 agents, including all of Group A itself,

as well has half of both High-Low biased groups (B1, which is biased toward R_1, and C1, which is biased towards R_2). Group C is the mirror image of Group B. Note that we have two subgroups, B2 and C2, whose reference networks are restricted to those who share their inherent utility High-Low distributions.[40]

As in other models, each group simply acts on its inherent utility in the first period. However, because we have multiple reference groups that employ different updating calculations, an order of updating was applied in all periods after 1; first Group A updated and acted, then B1, B2, C1 and finally C2. Thus each period has five mini-periods; when a group updates it considers the last move made by others in its reference group. This is by no means an entirely innocent stipulation: as W. Brian Arthur pointed out, in closely split populations those who move earlier can have significant effects on the outcome of convergence.[41] As some counterweight to the polar groups B and C, the random group, A, was thus given the first move in each period, giving random factors some advantage over the effects of High-Low polarity.

In the basic simulation, Group A had a modest 54% to 46% tilt towards R_1; Group B was High-Low polarized 60/40% towards R_1, while Group C was High-Low polarized 60/40% towards R_2 (actually, Group C2 was slightly more strongly polarized towards R_2, with 64% having High-Low bias towards R_2). Overall the entire population group of 150 was approximately split 51% R_1 to 49% R_2. As Figure 9 shows, convergence on R_1 occurred in ten periods. Interestingly, Group C2, which had a strong 64% High-Low bias towards R_2, and whose entire reference group also had a 60% bias toward R_2, began, as we would expect from our analysis in the last section, by moving *towards* R_2, at one point being 84% R_2 followers. Group C1 was, at it were, initially pulled in two directions: some of their reference group (A) was moving towards convergence on R_1, while C2 was moving toward R_2. For several periods, then, C1 remained unchanged, until the movement in Group A was strong enough to pull them towards R_1. And, in turn, that eventually pulled C2 in their wake, ending up with 100% R_1 convergence. The last to switch to R_1 were, as was expected, Quasi-Kantians favoring R_2 in Group C2 (and one Linear Agent).

[40] In these two subgroups with reference groups of 50 (B2, C2), all weighting systems were normalized so that maximum $n = 50$.

[41] W. Brian Arthur, *Increasing Returns and Path Dependence in the Economy* (Ann Arbor, MI: University of Michigan Press, 1994), see especially chap. 5

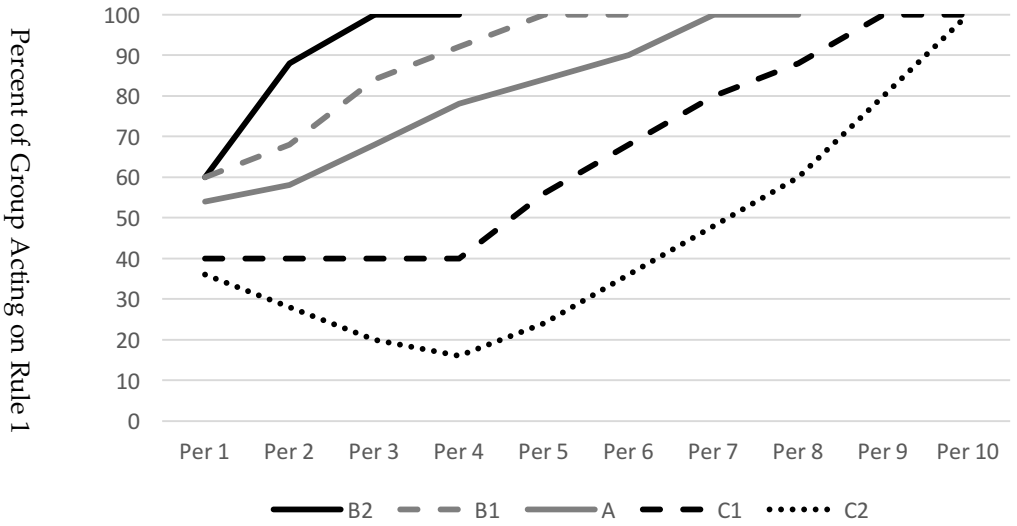

FIGURE 9: CONVERGENCE WITH FIVE DIFFERENTIAL REFERENCE GROUPS WITH HIGH-LOW BIOLARITY IN FOUR GROUPS

Diversity of types is important, though certainly not always necessary — if we have the right sort of type. In a simulation with the same distribution of inherent utilities as above, a homogeneous population of Linear Agents failed to converge on a rule in *any* of our five groups; the same was true with a pure population of Quasi-Kantians. In a homogeneous population of Highly Conditional Cooperators, however, convergence was quickly achieved, in five periods. This should not be surprising, since Highly Conditional Cooperators give great weight to high convergence. As I said, they are died-in-the wool Humeans.

In another simulation with all four types, Group A's inherent utilities were assigned randomly, resulting in 66% of Group A favoring R_1. Not surprisingly, full convergence was achieved in the population of 101 agents very quickly — in four periods. More interesting is what occurred when not only inherent utilities, but *agent types* were randomly distributed in the population. Here full convergence on R_1 was not achieved: 24 (of the 50) members of Group C (including all the High-Low biased members of C2) maintained a small R_2 network, while the rest on the population (126 agents) moved to R_1. In Group C1, those with High-Low evaluative utilities biased in favor of R_2 were almost all Linear Agents (with three Quasi-Kantians). As a result, those High-Low biased in favor of R_2 in C1 were not sensitive to movement in Group A to R_1, which in turn insulated C2 from the movement of the random group, A. Linear Agents engage in such gradual

updating that few could overcome their own the High-Low bias in favor of R_2, even though there was some movement within C1 to R_1.[42] Recall that in a homogenous population of Linear Agents none of the five groups achieved convergence. Again, we see how a diversity of types can generate agreement.

5 MODELING MORAL CHOICE

My aim has been to sketch some rather basic models of perfectly moral and rational agents, and to explore some dynamics of rational action that lead a population with somewhat stark moral disagreements to converge on a shared moral rule. Like social contract theories such as Rawls's, the point of our models is to understand rational moral persons and their choices. I have tried to show here that under some conditions characterized by deep diversity they would be able to organize themselves into freely-endorsed moral systems. Central moral controllers (such as social contract theorists) are not necessary. A number of parameters are relevant to these models: agents' information about others, the depth and extent of their moral disputes, weighting functions, sizes of, and links among, reference groups and so on. We have investigated some of these in a very preliminary way.

What is not a usefully manipulated parameter is whether the agents actually have moral motivations. Moral theory — at least public reason inclined theories — are interested in whether free and equal good willed persons with sharp differences can endorse common moral orders. We do not ask the Rawlsian contract theorist what the contract would be if some individual had a gun in the original position, though the presence of guns may certainly effect the institutionalization of the contract. Questions of non-moral motivation and actual power asymmetries become relevant when evaluating to what extent actual moral systems conform to our rather austere normative choice-based models. Here one might ask of self-organizing models, "How manipulable are they by an agent such as Charlie, who strategically misrepresents his moral utilities, and seeks to exercise power over others, to get others to follow his moral preferences?"[43] This leads to great complexities. How large of a system is Charlie in? What is his reference group, and who focuses on Charlie as critical to their reference groups? What are the views of others? Does Charlie organize with like-minded others to form a

[42] In another simulation of this population Quasi-Kantians (again, with the random distribution of agent-types), the entire population of 150 quickly converged — almost completely in four periods, with the last two in C2 adopting R_1 in Period 5.

[43] In modeling of how norm systems actually change this is a critical question. See Bicchieri, *Norms in the Wild*, chap. 5.

vanguard elite? Does he simply seek to manipulate through strategic "moral action" or does he undertake other, more overt, forms of persuasion (say, with guns)? One tiny step to examine these types of questions was made. A common worry is that a small, devoted, group of like-minded agents totally devoted to rule R_1 who, strategically, refuse to act on any other rule regardless of their true evaluation of it, could be decisive in pushing an entire group to R_1 when R_2 is actually favored by a very small majority. If the population is knife-edged but ever-so-slightly inclining toward R_2, perhaps our strategic R_1 supporters could reliably tip the group the other way. To take a first look at this possibility our earlier population of 101 with a knife-edged split of 51 R_2- and 50 R_1-inclining agents (Figure 5) was altered: 10% percent (5 agents) in the R_1-inclining group were removed (one from each weighting type, plus a randomly selected fifth) and replaced with purely strategic R_1 agents who simply act on R_1 no matter what others do. Ten percent seemed a reasonable number for a strategic R_1-party trying to manipulate the system. We might hypothesize that these agents would either tilt the convergence dynamic to R_1, or at least significantly slow down the process converging on R_2 (the result in Figure 5). In fact, they had no effect on the behavior of others: convergence of the rest of the population (96 agents) on R_2 occurred in the same number of periods (five), with the only difference being that the strategic agents were stuck in their own, five-member R_1 network. This gives us a very preliminary indication that a small (but not tiny) number of dedicated like-minded individuals cannot always easily manipulate, simply through their choices, even a closely split system.

6 MORAL FREEDOM AND UNITY IN DIVERSE, COMPLEX SOCIETIES

Contemporary moral theory and social philosophy divides into two opposing lines of thought or, we might say, research projects. The traditional, still dominant, project carries on with articulations of the "I believe we ought" view. These accounts are often sophisticated and admirable exercises in philosophical reasoning, building on the fundamental intuition that the best moral conclusion for one is the best for all. This research project has great difficulty in even making sense of the idea of moral diversity.[44] To be sure, there is moral disagreement and conflict — some pig-headed and ill-informed and some, perhaps, more reasonable — but the study of ethical life need be no more focused on diversity than is

[44] I have greatly benefitted from discussing the issues raised in this section with Piper Bringhurst.

physics.[45] The second line of inquiry seeks, in disparate ways, to make sense of the idea of fundamental moral difference. To Isaiah Berlin, the Romantic philosophers of the seventeenth and eighteenth centuries have "permanently shaken the faith in universal, objective truth, in matters of conduct" by showing that "ends recognized as fully human are at the same time ultimate and mutually incompatible."[46] Berlin repeatedly insisted that the Romantics taught us that there are many values, and that they are incommensurable; "the whole notion of plurality, of inexhaustibility, of the imperfection of all human answers and arrangements, the notion that there is no single answer which claims to be perfect and true ... all this we owe to the romantics."[47] I have tried to show here that this second line of inquiry — which somehow takes the diversity of moral conclusions as a basic datum of ethical inquiry — is fundamental to the social contract tradition. Once such diversity is understood not as moral reasoning gone awry, but as the crux of free human moral reasoning, moral diversity in some guise becomes the core of moral theory and social philosophy.

Yet the social contract — and this most definitely includes Rawls — never really advanced beyond what we might call "designs to manage difference."[48] Moral difference is seen as the fundamental problem for moral theory, but the aim is to plan for mediation via a social contract that rises above difference to show an underlying homogeneity. Certainly mediation and reconciliation are fundamental concerns of moral theory, for as soon as we cannot suppose that good moral reasoning alone shows us the path to a cooperative social life, we need to find new paths, which produce some unity in expectations and understandings *out of diversity*. F. A. Hayek stressed throughout his career that rational constructivism and planning is not usually a viable way to cope with heterogeneity. Central

[45] This, of course, can be a two-edged comparison. See Thomas Kuhn, *The Essential Tension* (Chicago: University of Chicago Press, 1977) and D'Agostino, *Naturalizing Epistemology* (London: Pelgrave, 2010).

[46] Berlin, "The Apotheosis of the Romantic Will" in *The Crooked Timber of Humanity: Chapters in the History of Ideas,* edited by Henry Hardy (Princeton: Princeton University Press, 1990): 207-37 at p. 237. See also Berlin, *The Roots of Romanticism* (Princeton: Princeton University Press, 1999), pp. 34ff; Bernard Williams, "Conflict of Values" in his *Moral Luck* (Cambridge: Cambridge University Press, 1981), pp. 71-82.

[47] Berlin, *The Roots of Romanticism*, p. 146. Rawls was sufficiently schooled in the first line of inquiry that he draw back from Berlin's talk of "competing truths" in morality, seeking to put questions of truth aside and focus on the concept of the reasonable. See Rawls, "The Independence of Moral Theory."

[48] An important exception to this broad claim is Muldoon's social contract in *Social Contract Theory for a Diverse World*. If we omit his commitment to the bargaining approach, his analysis has strong affinities to the view advanced here.

planners, be they economists or moral philosophers, do not have access to the necessary information about diversity: they can only cope with it by limiting admissible diversity, relying on "normalizing" assumptions about agents.[49] I have tried to take some small steps here in theorizing how we might think of moral theory without that form of central planning practiced by social contract theorists. The guiding idea is to model morally autonomous diverse agents making choices in the context of each other's choices, seeing what dynamics lead to a shared rule that all endorse. The motto of this project is that morality is best understood as a bottom-up affair. "The moral law is not imposed from above or derived from well-reasoned principles" but arises from the values of individuals and their distinctive searches for integrity and reconciliation in their social-moral lives.[50]

Philosophy & Center for the Philosophy of Freedom
University of Arizona

[49] See my *Tyranny of the Ideal*, esp. chap. 4.

[50] Frans de Waal, *The Bonobo and the Atheist* (New York: W.W. Norton, 2013), p. 228. This essay's epigraph was from page 23.

Appendix: Agent Types

N =	MC	LA	QK	HCC
1	0	0	0	0
2	0	0.01	0	0
3	0	0.03	0	0
4	0	0.03	0	0
5	0	0.04	0	0
6	0	0.05	0	0
7	0	0.06	0	0
8	0	0.07	0	0
9	0	0.08	0	0
10	0	0.09	0.02	0
11	0	0.1	0.04	0
12	0	0.11	0.06	0
13	0	0.12	0.08	0
14	0	0.13	0.11	0
15	0	0.14	0.13	0
16	0	0.15	0.15	0
17	0	0.16	0.17	0
18	0	0.17	0.2	0
19	0	0.18	0.24	0
20	0	0.19	0.29	0
21	0	0.2	0.34	0
22	0	0.21	0.39	0
23	0	0.22	0.42	0
24	0	0.23	0.45	0
25	0	0.24	0.47	0
26	0	0.25	0.51	0
27	0	0.26	0.53	0
28	0	0.27	0.55	0
29	0	0.28	0.58	0
30	0	0.29	0.61	0
31	0.01	0.3	0.62	0
32	0.06	0.31	0.63	0
33	0.11	0.32	0.65	0
34	0.15	0.33	0.67	0
35	0.19	0.34	0.69	0.01
36	0.21	0.35	0.71	0.015
37	0.23	0.36	0.72	0.017
38	0.24	0.37	0.73	0.0175
39	0.27	0.38	0.75	0.018
40	0.31	0.39	0.77	0.0185
41	0.33	0.4	0.79	0.019
42	0.35	0.41	0.81	0.0195
43	0.38	0.42	0.83	0.02
44	0.41	0.43	0.85	0.03
45	0.45	0.44	0.87	0.035
46	0.48	0.45	0.89	0.04
47	0.51	0.46	0.91	0.045
48	0.54	0.47	0.94	0.05
49	0.57	0.48	0.97	0.055
50	0.6	0.49	1	0.06
51	0.62	0.5	1	0.065
52	0.64	0.51	1	0.07
53	0.66	0.52	1	0.075
54	0.68	0.53	1	0.085
55	0.7	0.54	1	0.095
56	0.72	0.55	1	0.09
57	0.74	0.56	1	0.095
58	0.76	0.57	1	0.1
59	0.78	0.58	1	0.11
60	0.8	0.59	1	0.12
61	0.81	0.6	1	0.13
62	0.82	0.61	1	0.14
63	0.83	0.62	1	0.15
64	0.84	0.63	1	0.16
65	0.85	0.64	1	0.17
66	0.86	0.65	1	0.18
67	0.87	0.66	1	0.19
68	0.88	0.67	1	0.2
69	0.89	0.68	1	0.21
70	0.9	0.69	1	0.23
71	0.91	0.7	1	0.25
72	0.92	0.71	1	0.27
73	0.93	0.72	1	0.29
74	0.94	0.73	1	0.31
75	0.95	0.74	1	0.33
76	0.96	0.75	1	0.35
77	0.97	0.76	1	0.37
78	0.98	0.77	1	0.39
79	0.99	0.78	1	0.41
80	1	0.79	1	0.43
81	1	0.8	1	0.45
82	1	0.81	1	0.47
83	1	0.82	1	0.49
84	1	0.83	1	0.51
85	1	0.84	1	0.53
86	1	0.85	1	0.56
87	1	0.86	1	0.59
88	1	0.87	1	0.62
89	1	0.88	1	0.63
90	1	0.89	1	0.66
91	1	0.9	1	0.69
92	1	0.91	1	0.72
93	1	0.92	1	0.75
94	1	0.93	1	0.78
95	1	0.94	1	0.81
96	1	0.95	1	0.85
97	1	0.96	1	0.88
98	1	0.97	1	0.91
99	1	0.98	1	0.94
100	1	0.99	1	0.97
101	1	1	1	1

Key:

MC: Moderately Conditional

LA: Linear Agents

QK: Quasi-Kantians

HCC: Highly Conditional Cooperators