

A Puzzle for Contractualism

Joseph Heath
Centre for Ethics
University of Toronto
[draft]

One of the central and most attractive features of contemporary social contract theory is the idea that principles of justice exist in order to divide up the “benefits and burdens of cooperation.” There are many circumstances in which individuals are able to engage in mutually beneficial interaction, but on the condition that each exercise some restraint in the pursuit of his or her individual interest. Thus the situation calls for a measure of voluntary self-restraint, which each individual must (by and large and in general) be persuaded to undertake. The structure of the interaction, however, underdetermines the choice problem, in the sense that there are many different cooperative arrangements, each of which involves a different allocation of the burdens and benefits, but *all* of which are mutually beneficial. Thus a set of principles is required, in order to specify the precise modalities of cooperation – who does what, who gets what, who decides what, etc. – in a way that will be acceptable to all.

The central contractualist idea – articulated at its highest level of generality – is that the principles of justice specify the terms under which individuals would voluntarily *agree* to undertake a *cooperative* interaction. Both ideas are important: the fact that each individual must be induced to agree is what accounts for the deontological flavour of these principles (i.e. the fact that they do not always recommend maximizing aggregate welfare); while the fact that the interaction is cooperative explains why individuals might nevertheless be willing to accept something less than their maximal claim. This analysis, however, leaves open an important ambiguity with respect to the level at which these principles should be taken to apply. Perhaps the most natural way to apply them is at the level of particular interactions, such as business partners trying to set up a joint venture, siblings trying to divide up an estate, or shipwreck survivors allocating resources on a desert island. These people can benefit by cooperating with one another, yet failure to agree upon specific modalities has the potential to erode these gains. Thus a set of principles that can command convergence, with regard to the specification of these modalities, has obvious appeal. Furthermore, applying the principles of justice at this level is responsive to our everyday sense that these problems must each be solved in a way that is fair to the individual participants, e.g. that the partners, siblings, or survivors, should each be treated fairly with respect to this *particular* interaction.

I refer to this way of applying the central contractualist idea as microcontractualism, on the grounds that it treats the principles of justice as constraining individual conduct at the action-theoretic level. It has obvious appeal, but is also subject to certain powerful objections. In particular, this way of applying contractualist principles seems to bear more than a passing resemblance to libertarianism (indeed, David Gauthier's contractarianism, which is a paradigm instance of a microcontractualist analysis, is sometimes lumped together into the broader family of libertarian theories¹). There is an important difference, which is that libertarianism typically takes whatever the parties happen to agree to with respect to the division of benefits and burdens as authoritative, whereas contractarians – including Gauthier – instead favour the division that would be normatively prescribed, through the application of a set of principles that reflect the choices individuals would make, under more-or-less idealized conditions. There is, however, one important similarity, which is that both libertarianism and microcontractualism leave individuals without redistributive obligations toward those with whom they choose not to cooperate. Of course, while both Robert Nozick and Gauthier view this as an attraction of their respective views, many other philosophers regard it as a *reductio* (illustrating, if nothing else, the unhelpfulness of many of the moral “intuitions” that are routinely appealed to in contemporary moral philosophy). There is, however, a more firmly specifiable worry, which is that this feature of the view makes it possible for individuals to game the principles of justice, allowing them to achieve outcomes through selective association that could not be achieved within the scope of an ordinary cooperative interaction. This violates a stability property that is plausibly regarded as an important desideratum for any theory of justice.²

The most common response to this problem has been to shift the level at which the principles are applied, so that instead of being used to resolve the modalities of particular cooperative interactions, they are instead applied to “social institutions” more generally, and in the extreme, to “society” as a whole. This is, of course, the way that the classical social contract theorists conceived of

1 Most influentially, by Will Kymlicka, *Contemporary Political Philosophy*, 2nd edn. (Oxford: Oxford University Press, 2002), p. 128-138. Kymlicka does not actually explain why he considers the view libertarian, and says nothing about its strongly egalitarian features (such as Gauthier's claim that the state may rightfully confiscate the portion of Wilt Chamberlain's earnings that constitutes economic “rent,” – which is to say, almost all of it. See David Gauthier, *Morals by Agreement* [Oxford: Clarendon, 1986], p. 273). The egalitarian aspect is somewhat concealed by Gauthier's unorthodox, and ultimately unsuccessful, derivation of the minimax relative concession principle. If one looks instead at the equivalent result in bargaining theory, the Kalai-Smorodinsky solution, one can see that it incorporates an egalitarian “symmetry” axiom.

2 For the general flavour of this, see John Harsanyi, “A Simplified Bargaining Model for the n-Person Cooperative Game,” *International Economic Review*, 4 (1963): 194-220; Terje Lensberg, “Stability and the Nash Solution,” *Journal of Economic Theory*, 45 (1988): 330-341. The idea is that in an *n*-person solution, having one person exit with his just share should not cause all the remaining participants to want to renegotiate their shares.

the doctrine – as providing the terms governing the “civil condition” as a whole – and it is echoed in John Rawls's famous description of society as a “cooperative venture for mutual advantage.”³ I refer to this as macrocontractualism, for obvious reasons. Shifting to this level of analysis makes the social contract much more metaphorical, which has certain disadvantages, but it also helps to minimize the problem of selective association. Individuals cannot evade their obligations toward others simply by avoiding any sort of cooperative interaction with them; by virtue of belonging to the same society, they are part of a generalized system of cooperation, and so are subject to certain obligations that apply to everyone. Of course, the problem may still recur at the boundary, if one stops short of treating all of humanity as party to the contract. Rawls, for instance, treats “the” system of cooperation as secured by the basic structure of society, largely coinciding with the institutions of the nation-state. This, in turn, generates a set of reduced obligations towards foreigners, a position that has attracted a certain measure of resistance among those who feel that it represents a pinched, perhaps even ungenerous, response to the human condition.

In this paper, I would like to focus on a different, less well-known problem with macrocontractualism. By shifting the analysis up to the level of “society as a whole,” it is easy to lose track of the fact that individuals also expect their particular interactions and private associations to be fair, above and beyond whatever contribution the outcome may make to the fairness of the entire society (so that, for instance, even in the context of an unjust society, particular institutions or domains of interaction might nevertheless be just). Our sense is also that a just society is one in which all component institutions and associations – families, schools, churches, contracts, wills, corporations, etc. – can also be deemed to be just, in a relatively self-standing fashion. Yet macrocontractualism seems to lack the resources needed to ensure this. Indeed, contractualists such as Rawls wind up adopting a surprisingly laissez-faire standard to judge the particular cooperative projects that individuals may undertake (again, not far off from libertarian views). Hence the puzzle that emerges: if we start out at the bottom, with a micro perspective, insisting that particular cooperative interactions and small-scale institutions be internally just, then we have no assurance that this will all add up to a “just society” at the macro level. If, on the other hand, we start out at the macro level, and insist that society as a whole be just, then we lose the ability to insist that small-scale institutions and interactions be internally just.

My objective in this paper will be to map out in slightly greater detail how this puzzle comes

3 John Rawls, *A Theory of Justice*, rev. edn (Cambridge, MA: Harvard University Press, 1999), p. 4.

about, then suggest a strategy for resolving it that takes us beyond the standard flavors of contractualism. This involves adopting a cultural-evolutionary perspective, then interpreting the principles of justice in terms of a set of pragmatic-structural biases in the transmission and reproduction of social norms.

I

I would like to begin by providing an outline of what I refer to, perhaps tendentiously, as “minimally controversial contractualism,” then show how the puzzle follows quite immediately from it. The central advantage of contractualism, from the standpoint of its adherents, is the leverage it provides in responding to various forms of moral skepticism. First and foremost, contractualists are inclined to treat rational egoism as a serious concern, and therefore to worry about what Christine Korsgaard calls “motivational skepticism.”⁴ Even if we had the power to discern moral facts, or had access to clear and distinct moral intuitions, contractualists worry about how individuals are to be persuaded to respect these judgments in practice, especially when the moral rules demand that we set aside our self-interest – often in rather dramatic ways – in favour of the good of others. Contractualists generally would like to have something to *say* to the person who is unmoved by altruistic appeals. The contractualist approach takes as its point of departure the observation that adherence to moral rules typically produces benefits for others that are greater than the losses to the individual. Thus when *jointly* adopted, they produce mutual benefit, which is to say, they establish a system of cooperation. Morality involves sacrifice, but it also produces reward. Thus for those who worry about motivational questions, the focus on these rewards provides, if not a reason to act morally, at least a rationale for the way that morality constrains individual self-interest.

I mention this because, among critics, the contractualist focus on cooperation – and thus on mutual benefit – is often portrayed in a negative light, as though it were motivated by a desire to limit the scope of our obligations.⁵ Invidious comparisons are drawn to the expansive, almost promiscuous extension of moral duty among utilitarians, many of whom believe that we have unbounded obligations to improve the happiness of all living things, or luck egalitarians, who believe that we are literally responsible for bringing about the Kantian *summum bonum* (a task that Kant himself believed could only be plausibly undertaken by an omnipotent God, and even then would require an eternity to

4 Christine Korsgaard, “Skepticism about Practical Reason,” in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996): 311-334.

5 Brian Barry, *Theories of Justice* (Berkeley: University of California, 1989), p. 163.

achieve). And yet the problem with these sorts of high-minded ideals is that they tend to lack motivational efficacy. They overrule considerations of self-interest so completely that it becomes a serious question whether any of us could ever be justified in, say, buying a cup of coffee, much less a pastry to go with it. Because they show such wanton disregard for the interests of the individual, these views make it difficult to see how one could convince a person not already in their grip to take them seriously. They are, as it were, all stick and no carrot.

Contractualists are generally willing to sacrifice some of this loftiness, in the interests of producing norms that are more likely to have motivational efficacy. Rawls articulated this ambition quite clearly when discussing what he called the “strains of commitment.” Agreements that require us “to accept the greater advantages of others as a sufficient reason for lower expectations over the whole course of our life,” make what he describes as an “extreme demand” on individuals, compliance with which may well “exceed the capacity of human nature.”⁶ He took this as grounds for limiting the range of fair outcomes to those belonging to what John Nash referred to as the “feasible set” in cooperative interactions.⁷ The important point is that the focus on cooperation, and hence the willingness to limit the scope of our obligations both with respect to the persons to whom they are owed and the benefits that are subject to their claims, is not always – or even usually – motivated by a desire to minimize the claims that others can make on us. It is more often a consequence of a genuine concern about motivational skepticism. Mutual advantage seems to provide a happy *via media* between the vulgar appeal to self-interest and the question-begging reliance upon existing moral commitments.

With respect to the *content* of our moral judgments, contractualists are also worried about skepticism, particularly when it comes to the principle of equality. While some philosophers seem content simply to posit a commitment to equality, as some sort of ultimate value, not susceptible to further justification⁸, contractualism recommends itself to those who would like to have something to say to those who don't already share this commitment (or worse, who actively distance themselves from it, on the grounds that it is nothing but a rationalization of envy). It is important to recognize that equality is an extremely demanding moral ideal, one that can impose onerous obligations upon the individual. Furthermore, since any serious commitment to equality must involve a willingness, at times, to level down, equality can enter into tension with other, quite plausible, moral ideals.⁹ All of this

6 Rawls, *Theory of Justice*, pp. 177-8. He continues: “In fact, when society is conceived as a system of cooperation designed to advance the good of its members, it seems quite incredible that some citizens should be expected, on the basis of political principles, to accept lower prospects of life for the sake of others.”

7 John Nash, “The Bargaining Problem,” *Econometrica*, 18 (1950): 155-162.

8 E.g., see G.A. Cohen, *Rescuing Justice and Equality* (Cambridge, MA: Harvard University Press, 2008), p. 7.

9 For example, what Larry Temkin calls “the Slogan” strikes most people as being extremely plausible at first glance. See

generates a significant burden of justification, particularly toward those who can be expected to be the losers in any egalitarian redistribution.

In response to this challenge, contractualists have a rather simple and very powerful claim. Starting with Hobbes, the fundamental argument for the principle of equality has been that it arises out of a symmetry condition that must be satisfied in order to secure agreement. Whether it is “splitting the difference,” or creating equal shares, or flipping a coin to choose the winner, everyone is familiar with the way that equalization can be used as a technique to minimize, and often eliminate, a particular sort of objection to a cooperative enterprise. Thus equality is not simply posited, as an ultimate value, it is derived from the constraints that must be satisfied in order to achieve agreement. To the “losers” in any egalitarian distribution, then, who want to know why they should be the losers, one can point out that the losses are entirely hypothetical. Without an ongoing system of cooperation, there would be nothing to lose. Yet the cooperative scheme is made possible only by the willingness of all participants to play along, a willingness that is, in turn, brought about by the fact that the benefits and burdens of the system are distributed and borne equally.

The standard way of illustrating this is with a prisoner's dilemma, as shown in Figure 1. Note that the numbers need not be taken to represent utility, but could be anything that the two players are able to produce through cooperation, whether it be increased life-expectancy, calories available for consumption, travel speed, evolutionary fitness, or, of course, preference-satisfaction. The set of possible payoffs, obtainable through either randomization or repeated interaction, is shown as the diamond-shaped region in panel B. Since each individual is able to guarantee herself a payoff of 1 through straightforward maximization, the space of mutually beneficial arrangements is the set of points to the north-east of (1,1) – the “feasible set,” or what Gauthier described as the potential “cooperative surplus.” The fact that there is a continuum of possible cooperative arrangements reveals the extent to which utility-maximization (or self-interest, narrowly construed) underdetermines the interaction.¹⁰ There are literally an infinite number of possible cooperative arrangements, even within the scope of this highly simplified model of interaction. Furthermore, if one were to imagine a simple “alternating offers” bargaining model (where one player proposes a particular cooperative arrangement, with the other having a choice of either accepting it or proposing a counteroffer) absent some penalty for delay, and among players animated only by self-interest, it is easy to see that the game will go on

his *Inequality* (Oxford: Oxford University Press, 1993), pp. 248-255. Explaining what is wrong with it requires considerable subtlety.

¹⁰ This is codified in the form of the folk theorem for repeated games. See Drew Fudenberg and Eric Maskin, “The Folk Theorem in Repeated Games with Discounting or with Imperfect Information,” *Econometrica*, 54 (1986): 533-554.

forever. The players will simply take turns making self-serving proposals, which the other will reject in favor of an equally self-serving counteroffer.

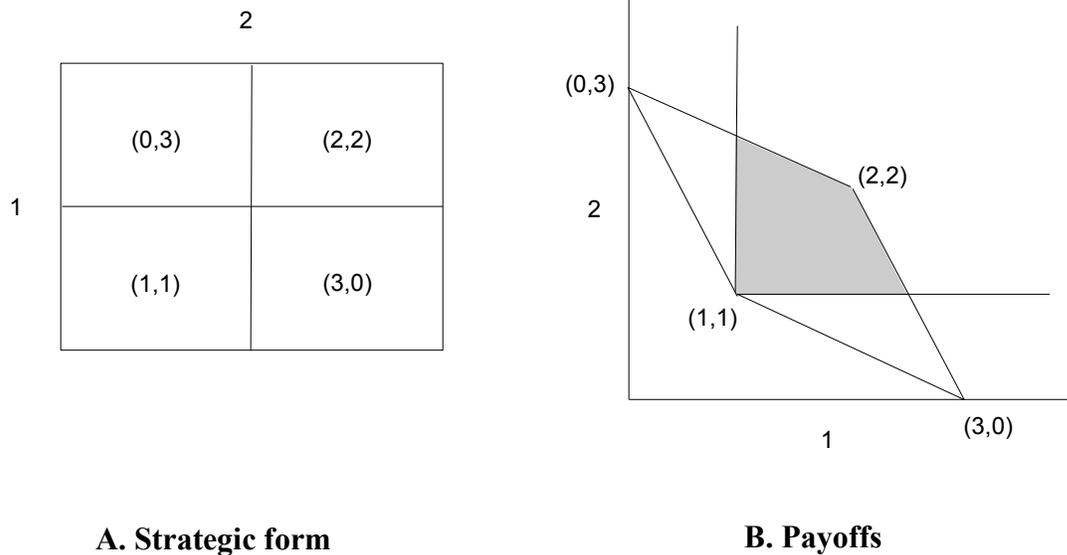


Figure 1. Prisoner's dilemma

Thus it is no great stretch to imagine that a set of normative principles is called for, in order to specify what should count as a reasonable agreement, and that the parties should accept such an agreement, not because it coincides with their self-interest, but precisely out of a recognition that self-interest fails to provide an acceptable basis for agreement, and that given this failure, the proposed principles are reasonable. Of course, there may be all sorts of “thick” cultural resources that the parties can appeal to in order to resolve such problems (such as an inherited set of social norms that specify how different sorts of interactions should be organized). Indeed, it has been observed that when “public goods” games are played in some non-Western societies, where the practice of “the psychology experiment” are unfamiliar, individuals often respond by searching for a “cultural template” for the interaction.¹¹ So after thinking about the structure of the interaction, they may say something like “this is just like in the village, when everyone contributes to repairing the road.”¹² They then act as they

¹¹ Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath, “In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies,” *American Economic Review* 91 (2001): 73–78.

¹² *Ibid.*, p. 76.

would if it were an interaction of that sort.

If one assumes, however, that thick resources of this type are unavailable – either because there is no cultural template, or because there are multiple templates and the choice of one is controversial – there are two principles that seem to be suggested by the very structure of the interaction, or that can be applied without drawing upon any particular cultural resources. First, if there is an arrangement that makes both individuals better off, then it would seem obviously superior to one in which they are both worse off. This judgment can be made without even getting into the details of the case. Articulating this idea as a principle yields the familiar Pareto-superiority criterion. Applying that principle to the problem at hand results in all points in the feasible set that are Pareto-inferior to some other point in that set being discarded as candidates for agreement. What remains are the set of Pareto-optimal points, shown in Figure 2. It is worth repeating the familiar observation that the ordering of points imposed by the Pareto-superiority criterion is incomplete, since the set of Pareto-optimal points are unranked vis-a-vis one another. This means that the set of points shown in Figure 2 can be thought of as an indifference curve – each outcome in the set is just as good as any other, from the standpoint of efficiency.

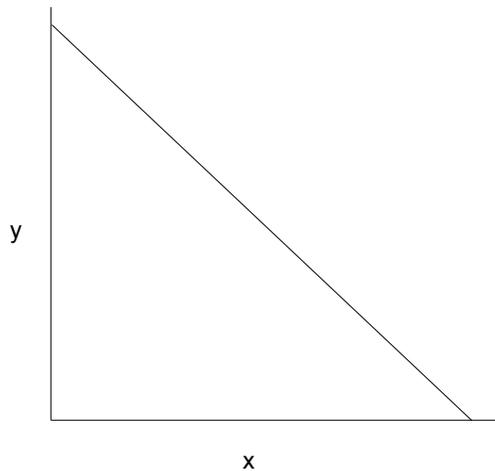


Figure 2. The set of Pareto-optimal outcomes

The second principle is less self-evident, but will be familiar to anyone who has spent some time dealing with children. One obvious way of minimizing objections to a proposed distribution is to avoid giving anyone an incentive to switch places, or allocations, with anyone else. In a welfarist framework this generates a symmetry (or anonymity) principle (that no player should want to switch places with another player); in a resourcist framework it generates the envy-freeness principle (that no player should want to acquire the allocation of any other player). Either way, it suggests that the set of points in the feasible set that generate a desire to switch places on the part of any player should be discarded. Call this the egalitarian principle. What remains after its application is the set of symmetric, or envy-free points, shown in Figure 3. Here it is worth making the less-common observation that the ordering of points imposed by this principle is also incomplete, in a way that is precisely complementary to that of the Pareto principle. Thus Figure 3 again can be thought of as a social indifference curve – each outcome in the set is just as good as any other, from the standpoint of equality.

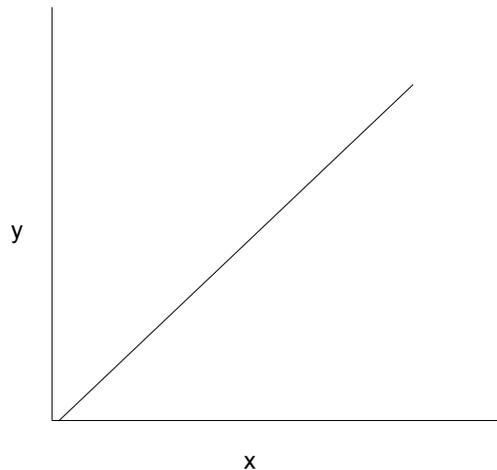


Figure 3. The set of symmetric, or envy-free outcomes

In the simplest of cases, under a first-best scenario, the intersection set of these two curves will be a single point (sometimes known as the efficient equal allocation).¹³ Contractualism then selects that arrangement as the most reasonable, not because it has any intrinsic merit, but simply because it does not give rise to any of the obvious objections that every *other* point in the feasible set would give rise to. (If one thinks of this in terms of the Parfit-Scanlon “complaint” model, one can see that moving to an efficient allocation eliminates one type of complaint, while moving to an equal allocation eliminates another. Thus one can derive the two principles from the most minimal version of the complaint model, one that does not need to get into the dicey business of distinguishing “stronger” from “weaker” complaints.)

In the real world, however, situations may easily arise in which it is possible to make significant improvements with respect to one of the two principles, but only by accepting an arrangement that is worse with respect to the other. Thus the question arises how much inequality one should be willing to accept in order to achieve gains in efficiency, or how much inefficiency one should be willing to accept in order to get improvements in equality. One way of resolving this is to assume that the further the status quo is from the ideal, with respect to either principle, the more strenuously players will object to a deviation of a given magnitude from it. If one assumes that the most favored arrangement will then be the one that players object to least strenuously, the result is a prioritarian ordering, in which benefits to an individual can justify departures from equality, but where these benefits “count” for less, the further away one gets from this ideal. The most well-known formula from making this tradeoff is the Nash bargaining solution, which favors the arrangement that maximizes the *product* of the benefit received by each player. This is shown in Figure 4 as a social indifference curve N , which is contrasted with the utilitarian solution U (exhibiting complete indifference to the distribution of benefit between the two individuals), and the difference principle D (exhibiting complete indifference to the allocation of the better-off individual, so long as that person remains the better-off).

¹³ It is a single point only because the problem involves a single distribuendum, and is therefore one-dimensional. When extended to $n > 1$ goods (as in a typical ‘resourcist’ framework), the set of envy-free allocations typically becomes a n -dimensional space containing multiple Pareto-optima. Thus some additional resources must be introduced in order to pick out a single solution. See Joseph Heath, “Dworkin’s Auction,” *Politics, Philosophy and Economics*, 3 (2005): 313-335.

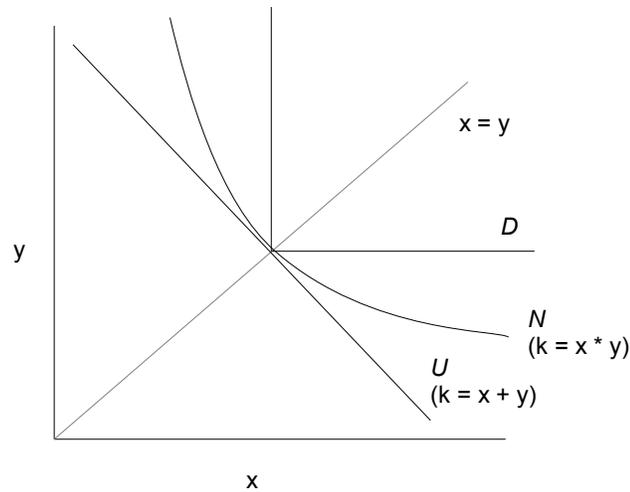


Figure 4. Social indifference curves

To see how the bargaining solution can be applied, consider a case such as Taurek’s numbers problem.¹⁴ There are five people on one island, one person on another: you have the opportunity to save the inhabitants of only one island. The maximizing solution (U) says you should save the five (and *ex ante*, with each individual having an equal probability of being on either island, this solution Pareto-dominates the alternatives). Yet the “fair” solution (D) would assign each individual an equal chance of being saved, which can be accomplished by flipping a coin, in order to decide whether to rescue the one or the five (this gives each person a probability of exactly $\frac{1}{2}$ of being saved). Now, suppose that one is torn between these two considerations. One doesn’t like the idea of imposing certain death upon the lone individual just because he had the bad luck of winding up on the wrong island, and yet one can’t avoid the feeling that letting five people die in order to save one is a terrible waste. What is needed is a way of balancing the two considerations against one another. The Nash Bargaining Solution does so: maximizing the product of each individual’s utility gain from the rescue suggests that you construct a weighted lottery that gives the lone individual a 1 in 6 chance of being rescued.¹⁵

¹⁴ John Taurek, “Should the Numbers Count?,” *Philosophy and Public Affairs*, 6 (1977): 293–316.

¹⁵ To see how: assign death a value of 0, life a value of 1, so that utility numbers are the same as the chances of living

(Furthermore, the solution automatically readjusts the lottery if one adds or subtracts people from either of the two islands.)

My intention is not to suggest that this solution is uncontroversial (although, if one had a rescue team that was composed of half consequentialists and half deontologists, who could not agree what to do, and so one wanted to split the difference between them, this would not be a bad way to do it). My reason for describing it as minimally controversial is that it is formulated at a higher level of generality than any of specific solution concepts offered by Rawls, Gauthier, and other contractualists. Indeed, all of these theories can be regarded as instantiations of the more abstract two-principle schema – specifying the equisandum more precisely, developing a corresponding formulation of the two principles, specifying a mechanism for trading off equality against efficiency, and of course, specifying the level at which cooperation is to be conceived. What I have outlined is a more “generic” (in the narrowest etymological sense of that term) form of contractualism.

II

Suppose one were to accept something like this “minimally controversial contractualism,” and agree that it provides an adequate template for the specification of a set of “principles of justice.” The question then becomes, how do these principles become effective in everyday life? How does the “rational” become “real”?

Perhaps the most natural approach, in answering this question, is to build the principles directly into one's model of practical rationality, by assuming that agents in some way make use of these principles in deciding what to do. This is Gauthier's approach.¹⁶ Thus he argues that in cases where the strategic equilibrium of an interaction is Pareto-optimal, agents reason in accordance with the standard canons of rational choice theory, but when faced with a potentially suboptimal equilibrium, they switch gears and begin to cast about for a cooperative solution. Once the players have established that their interaction partner is likely to cooperate, they apply Gauthier's favored contractualist solution concept (“minimix relative concession”) to the feasible set of the anticipated interaction, then carry out the actions needed to bring about that outcome. When both parties do the same, they are able to coordinate on a cooperative solution. (One can find a similar architectonic in “strong reciprocity” models of the

assigned by the lottery. One wants to construct a lottery that assigns a chance p of being rescued to the lone individual, and therefore a $1-p$ chance of rescue to the five other people, such that the product $p * (1-p)^5$ is maximized. The utilitarian solution (0,1,1,1,1) has a product of 0, the “fair” solution (.5, .5, .5, .5, .5) has a product of .0156. The product is maximized at approximately: (.165, .835, .835, .835, .835).

¹⁶ This approach is shared by contractualists ranging from T.M. Scanlon to Ken Binmore. I will not be discussing either of these views here, simply because both theorists downplay the importance of cooperation.

evolution of cooperation.¹⁷⁾

This “microfoundational” approach is extremely intuitive, in part because almost every theory in normative ethics has this structure (i.e. morality is thought to have practical effect because people take moral considerations into account when deciding what to do). When combined with the contractualist emphasis on cooperation, however, it generates some perverse consequences. This is because the principles of justice, on this conception, constrain individuals' actions only within the scope of cooperative interactions. They tell you how you should treat a person with whom you are cooperating. They do not tell you, however, with whom you should be cooperating. In this respect, the use of a prisoner's dilemma, in particular a two-player prisoner's dilemma, as the central model in the development of the theory is extremely misleading. This is because in a two-player model, each player's participation is necessary to the execution of the cooperative scheme. With three players, however, it may be the case that two players can cooperate without needing to include the third, and so strategic considerations enter into the choice of interaction partner.

In order to accommodate this added layer of complexity, the solution needs to be formulated in the language of cooperative game theory. This is designed to model interactions in which players can not only act on the basis of individual strategies, but can also form coalitions that can act on the basis of joint (i.e. cooperative) strategies. In order for the outcome of a multi-player interaction to be stable, then, it must not only be the case that no individual has an incentive to defect, but that no coalition (i.e. proper subset of the total number of players) has an incentive to defect. An outcome that possesses this property is described as being in the core of a game. Many games, however, do not have a core – which means that for every outcome achievable by the group as a whole, there will always be some proper subset of the total set of players that could do better for each of its members by leaving the “grand coalition” and acting on its own.

To take the simplest example of this, consider a group of three adventurers who discover a treasure chest deep in the jungle. It takes two to carry the chest out, but only two. It is not difficult to imagine that, for any proposed arrangement in which the three adventurers take turns carrying the chest, then split the reward into $\frac{1}{3}$ shares, there is a more attractive arrangement in which just two of them carry it, then split the reward into $\frac{1}{2}$ shares (all it takes is for the value of the reward to significantly outweigh the disutility of carrying it). Thus while the three of them have the option of cooperating with one another as a group, they need not do so. It is possible for two of them to cooperate

¹⁷ See, most recently, Samuel Bowles and Herbert Gintis, *A Cooperative Species* (Princeton: Princeton University Press, 2011), pp. 20-21.

while excluding the third; indeed, it is advantageous for the two of them – any two – to do so. Thus the game has no core. This raises two problems: first, the resulting division of the treasure ($\frac{1}{2}$, $\frac{1}{2}$, 0) seems to be, as Gauthier puts it, “transparently unjust.” And second, the question of who gets to participate in the winning coalition and who doesn't seems open to being resolved in a completely unprincipled manner – two players might decide to cooperate simply because they like each other, or because they were born in the same town, or because they share what is referred to in civil rights law as “prohibited grounds for discrimination.”

While this example may seem fanciful, the structure of the interaction is actually quite common, simply because the expansion of any cooperative scheme beyond a certain size often generates diminishing returns, which may give members of that scheme an incentive to admit fewer than the total number of potential cooperators. This is the classic problem afflicting worker cooperatives, for instance, which is why they were often received with hostility by egalitarian socialists.¹⁸ Under a profit-sharing regime, worker coops have an incentive to bring in new members up until the point at which the *average* profit is no longer increasing. Thus they will hire fewer workers than capitalist firms, which continue to hire labour so long as the *marginal* contribution to profit is positive.¹⁹

Because of this, worker coops will stop bringing in new members at a point at which absolute profitability could still be increased by hiring more labour. This could be simply a deadweight loss associated with that organizational form. A deadweight loss, however, is nothing but an unrealized opportunity for cooperation. Thus there is an incentive for the co-op to expand production by creating a secondary class of worker, brought in on a fixed wage, on the same terms that they would be in a capitalist firm. (One can see this sort of an arrangement in the structure of a typical law firm, with its division between “partners” and “associates.”) This is, of course, also a cooperative scheme; the fact that net revenue is positive at the margin means that both the co-op members and contract workers benefit. The problem is that it creates two tiers of workers within the firm (i.e. it creates a situation where “some are more equal than others”). Furthermore, because the introduction of contract labor increases profitability without expanding the number of co-op members, it increases the profit-share of each member, and thereby encourages the charge that they are exploiting those who weren't lucky enough to get in “on the ground floor.” Many people – including, over the years, many socialists – have had the egalitarian intuition that workers in a firm doing the same job should have the same status, and be entitled to the same rewards. And yet the interaction here has no core. The members of the

18 E.g. see Beatrice Potter Webb, *The Co-operative Movement in Great Britain* (London: Swan Sonnenschein, 1891).

19 Benjamin Ward, “The Firm in Illyria: Market Syndicalism,” *American Economic Review*, 48 (1958): 566-589 at 578.

cooperative would rather not hire the contract workers at all than bring them in on equal terms.

The general problem here is that of determining what Gauthier calls the “appropriate cooperative infrastructure.”²⁰ If there is a potential cooperative interaction between several people, is it acceptable for them to form coalitions first, and then have the coalitions enter into a cooperative arrangement with one another? If so, then determining the proper modalities of cooperation will involve applying the principles of justice several times, to several distinct cooperative surpluses. The results of this will almost inevitably be different from those that would be obtained by applying the principles of justice just once to the grand coalition. Suppose that four players can generate a cooperative surplus of \$12, but a particular pair of them can generate a surplus of \$8. A straightforward egalitarian division among the grand coalition would produce an allocation of (\$3,\$3,\$3,\$3). If, however, the pair are allowed to form a coalition first and split the \$8 between themselves, then enter into an agreement with the remaining two to realize the additional \$4, a two-step egalitarian division will produce a final allocation of (\$5, \$5, \$1, \$1). There are times when this seems appropriate, even necessary, if we hope to make the application of the theory at all tractable. Consider the case of two firms entering into a contract. One is inclined to assess the fairness of the contract by examining the way that it divides up the benefits of the particular exchange that it facilitates, while ignoring the question of how each firm then engages in an internal division of the advantages it receives (and certainly without trying to level the advantages across members of the two firms). And yet in other cases, such as the worker's cooperative, the two-step application of the principles seems to create a loophole that individuals can use to achieve outcomes that are entirely contrary to the spirit of equality.

Gauthier admits, quite directly, that he has no solution to this problem.²¹ The natural temptation, of course, is to say that all this strategizing about who to cooperate with is foreign to the idea of justice. People should not be able to cherry-pick their interaction partners in such a way as to minimize their redistributive obligations. Thus the principles of justice should always be applied to the grand coalition. One should be obliged to treat as a partner in cooperation anyone with whom one can cooperate, without coalitions, partial agreements or side deals. It is easy to see, though, that this pushes the entire framework in the direction of macrocontractualism. After all, it suggests that one has obligations of

20 David Gauthier, “Moral Artifice,” *Canadian Journal of Philosophy*, 18 (1988): 386-418, at 397.

21 David Gauthier, “Fairness and Cores: A Comment on Laden,” *Philosophy and Public Affairs*, 22 (1993): 44-47 at 47. Anthony Laden was probably correct to point out that, with this framework, the biggest issue of justice become what sort of game the players wind up playing, not the particular outcome they receive (since the structure of the game is going to determine opportunities) – see Anthony Simon Laden, “Games Philosophers Play: A Reply to Gauthier,” *Philosophy and Public Affairs*, 22 (1993): 48-52. In the treasure-chest example, it is certainly true that the metagame, in which it is decided which of the two will be the ones to carry it out, is where the action occurs. But Gauthier was certainly right, as well, to point out that the strategic structure of our interactions is usually unchosen.

justice, not only toward those with whom one actually chooses to cooperate, but toward those with whom one merely *might* cooperate. This will certainly be everyone within a very large radius. So without even getting into the special problem of children, the handicapped, the sick and the elderly – those who may not be in a position to offer any cooperative benefits to anyone²² – there is already good reason to want to break with the action-theoretic, microcontractualist perspective, and to apply social contract principles to “society” as a whole.

The result is the familiar macrocontractualist framework, with its stylized representation of society as a “cooperative venture for mutual advantage,” and the theory of justice interpreted in terms of principles that would bring about agreement, not in the real circumstances of choice, but in some hypothetical scenario. This allows one to insist that the “basic structure” of such a society – the basic framework of law, the major social institutions that determine life chances, such as the education system and the labor market – treat everyone equally, without worrying too much about the fact that not everyone will be making a positive contribution to this cooperative project, and even among those who do, the nature and quality of the contribution made will vary enormously. This puts an end to all strategizing about interaction partners, by assuming that, for the purposes of determining entitlements and responsibilities, everyone can be assumed to cooperate with everyone else, and that differences in contribution will all come out in the wash, as it were, when generalized across society as a whole.

The downside of this construct is that it solves one problem at the expense of creating another. Even while endorsing the idea that the major set of institutions in our society should be just, we also tend to judge particular cooperative arrangements, not in terms of their contribution to the justice of society as a whole, but on a relatively self-contained basis, using the same set of principles. Strictly speaking, the macrocontractualist should not be doing so. In particular, if one truly believes that the principle of equality is derived from the contract thought-experiment, then one should not be using any sort of egalitarian intuitions when judging particular interactions or institutions. Rawls, it should be

22 Shifting to a macrocontractualist perspective allows one to solve this problem – what Peter Vanderschraaf calls the “vulnerability objection” – rather easily (“Justice as Mutual Advantage and the Vulnerable,” *Politics, Philosophy and Economics*, 10 [2011]: 119-147). Assuming diminishing returns to consumption, there are advantages to be had from a social system that permits individuals to shift in and out of contributory roles while maintaining some level of consumption. It follows quite immediately from the folk theorem that a system in which individuals share the benefits of their labour when they are active, and receive benefits when they are inactive, can be sustained as an equilibrium of a repeated game. The fact that some people may never become active need not undermine the equilibrium – again, for obvious folk-theorem reasons. For an intergenerational model with a similar structure – agents shifting in and out of contributory and non-contributory roles – see Joseph Heath, “Intergenerational Cooperation and Distributive Justice,” *Canadian Journal of Philosophy*, 27 (1997): 361-76. Critics of contractualism typically err in presupposing that “reciprocity” must involve *direct* reciprocity, and so fail to appreciate how flexible and robust systems of *indirect* reciprocity can be.

noted, was consistent in this regard, in that he refrained from any attempt to assess the fairness of particular interactions (or, in the later formulation, refrained from using “political” principles of justice to assess them). His acolytes have sometimes not paid enough attention to passages such as the following, from *A Theory of Justice*:

It is a mistake to focus attention on the varying relative positions of individuals and to require that every change, considered as a single transaction viewed in isolation, be in itself just. It is the arrangement of the basic structure which is to be judged, and judged from a general point of view. Unless we are prepared to criticize it from the standpoint of the relevant representative man in some particular position, we have no complaint against it. Thus the acceptance of the two principles [of justice] constitutes an understanding to discard as irrelevant as a matter of social justice much of the information and many of the complications of everyday life.²³

It seems clear the Rawls' macrocontractualism lacks the resources to say anything critical about the “two-tiered” workers's cooperative – indeed, this sort of inequality seems to be one of the “complications of everyday life” that must be discarded as irrelevant. First of all, Rawls makes it clear that corporations (and cooperatives) are outside the basic structure of society, since they are voluntary associations.²⁴ This means that their internal structure (and division of advantages) cannot be directly assessed as just or unjust. Second, the impact that the inequality between the two tiers of workers would have on inequality in society at large would be difficult to assess (and, to the extent that allowing cooperatives to hire on contract is likely to mitigate their otherwise lamentable tendency to generate unemployment, it might even generate benefits for the “worst-off representative individual”), and therefore qualifies as one of the factors that would be too complicated to assess. Thus the only criterion that seems available to assess this organizational structure is that of voluntariness and conformity to law – did it come about through an exercise of the rights and liberties that individuals are accorded by the first principle of justice, and in conformity with the relevant enabling legislation (i.e. corporate or cooperative law)? If so, then the outcome is a matter of pure procedural justice.

Thus it is not obvious that Rawls is in a position to say anything about the firm that differs in any significant respect from the libertarian-contractual view that one finds in, say, Frank Easterbrook

²³ Rawls, *A Theory of Justice*, pp. 87-88.

²⁴ *Ibid*, p. 126.

and Daniel Fischel.²⁵ As long as everything is clearly announced in advance, everyone freely accepts the terms, and everyone retains a right of exit, it would seem that there should be no limits on terms of employment that a firm can offer (or the structure of shares that it can issue). Needless to say, Rawlsians have often felt that they have quite a lot more to say about these issues. Setting aside those who simply apply the difference principle directly to the distribution of advantages within the firm (ignoring all of the reasons that one cannot do this²⁶), Rawlsians have tried all sorts of subtle strategies to derive constraints on the way that firms can treat their workers, shareholders, customers, etc. To take just one example among many, Nien-hê Hsieh has argued (in “Rawlsian Justice and Workplace Republicanism”) that the standard capitalist firm is organized on the basis of authority relations that raise troubling issues of social justice. He posits a “basic right to protection against arbitrary interference” as part of the basic structure of society, then tries to show that a right to exit from employment does not provide adequate protection of this right.²⁷ His argument, however, appeals to the *cost* that this sort of exit typically imposes upon workers, and claims that workers cannot reasonably be expected to shoulder this burden. “Reasonable” here means, of course, “unjust,” but that simply begs the question, since the idea that there is a conception of justice governing these relations that is in some sense egalitarian is precisely what needs to be shown.²⁸

One can see the problem crop up in other areas as well. It explains, presumably, Rawls’s extremely ambivalent attitude toward the inclusion of the family in the basic structure. While claiming that families are part of the basic structure (on the grounds that they are essential to the orderly reproduction of society as “a scheme of social cooperation over time”), he then goes on to describe them as associations that arise *within* that structure, which are not subject to “political” principles of justice. This is because he doesn’t want to see their “internal affairs” subject to principles of distributive justice such as the difference principle. Yet it seems obvious that there are certain family structures, particularly those involving gender inequality, that we are inclined to regard as unjust. And yet it is difficult to see how Rawls could apply any standard but respect for basic rights, voluntariness and right of exit.

25 Frank H. Easterbrook and Daniel R. Fischel, *The Economic Structure of Corporate Law* (Cambridge, MA: Harvard University Press, 1996).

26 See remarks in John Rawls, “Idea of Public Reason Revisited,” *University of Chicago Law Review*, 64 (1997): 765-807 at 789-90.

27 Nien-hê Hsieh, “Rawlsian Justice and Workplace Republicanism,” *Social Theory and Practice*, 31 (2005): 115-142 at 128. Hsieh’s argument is considerably more subtle than the standard “Rawlsian” approach, which simply ignores the fact that corporations are not a part of the basic structure.

28 In later work, Rawls states that the internal affairs of associations must be governed by “some conception of justice (or fairness),” just not a “political” conception (“Idea of Public Reason Revisited,” p. 790).

The puzzle for contractualism, therefore, stated at its highest level of generality, is simply that we are inclined to apply principles of justice – particularly conceptions of equality – at both the micro and the macro level simultaneously. On the one hand, contractualists are like most people, in that they tend to worry about the overall distribution of advantages in society at a very high level of abstraction. Thus it is widely thought that the GINI coefficient, or the poverty rate, or gender inequality in various occupational spheres, reveal something important about the “justice” of the institutional arrangements of a society. At the same time, we tend to judge particular interactions and allocation rules according to primarily internal factors, without looking at total endowments, or their impact on the distribution of advantages in society at large. With education, for example, the baseline seems to be that of equal funding per pupil, even though this interacts with differences in natural ability in a way that produces highly unequal outcomes in several different dimensions.²⁹ Resource allocation in health care, although exhibiting a strong concern for equal treatment, is restricted almost entirely to medical criteria, without reference to broader circumstances of patients. The treatment of minority shareholders in corporate law is subject to a variety of restrictions that exhibit a concern for fairness, but which are applied without reference to who these shareholders are, what other investments they hold, and so on. As Jon Elster has observed, “Those who are entrusted with the task of allocating a scarce good rarely if ever evaluate recipients in the light of their past successes or failures in receiving other goods. Local justice is largely noncompensatory. There is no mechanism of redress across allocative spheres.”³⁰

Here is the problem: people have a fairly standard set of ideas about fairness, which we apply both to particular interactions, to medium-sized institutions taken singly, and to “society as a whole.” But it is not clear how one can consistently move from one level to another. There is a compositional fallacy in thinking that if you guarantee that the distribution of the cooperative surplus conforms to a set of principles of justice at the lowest level of individual interactions, adding up the results of these interactions will produce a distribution of the aggregate cooperative surplus (i.e. at the level of society as a whole) that conforms to these same principles. (So, for instance, just because every cooperative interaction between men and women is one that respects principles of gender equality, it does not follow that society as a whole will exhibit what is conventionally thought of as “gender equality.” It

29 It should be noted that luck egalitarians sometimes forget that they are not committed to equal funding in such areas, but rather to what Philippe van Parijs helpfully describes as “differential transfers, in amounts inversely related to people’s level of talent,” *Real Freedom for All* (Oxford: Clarendon, 1995), p. 61. Eric Rakowski, for example, after articulating a forceful commitment to luck egalitarianism, goes on to defend a scheme that would merely immunize incomes against differences in natural endowment, e.g. “guaranteeing that all who did the same work would receive the same income, as would be the case in a world where talents were equal and markets perfect.” *Equal Justice* (Oxford: Clarendon, 1991), p. 144.

30 Elster, *Local Justice* (New York: Russell Sage, 1995), p. 133.

depends also on the *pattern of association* that prevails between men and women.) The flip side of the coin is that guaranteeing that society as a whole respects certain principles of justice provides no assurances that this will percolate down successfully and produce interactions at the lower levels that respect anything like the same principles of fairness.

I describe this as a “puzzle” and not an “antinomy” because the problem can easily be resolved by anyone willing to bite the bullet, and simply apply the principles of justice at a particular level, damn the consequences at the other. Microcontractualists can adhere consistently to their view by rejecting any sort of “patterned” conception of justice at the macro level. Macrocontractualists can adhere consistently to their view by embracing voluntarism as the primary standard of rightness at the interactionist level. It is only if one wants to judge things at both levels using at least similar principles that there is a problem.

III

Before wrapping things up, I would like indicate briefly one of the directions that contractualists might go, in order to find a solution to this difficulty. I do so not because I have any proper solution worked out, but simply because I would like to offer some resistance to the almost inevitable temptation to avoid the problem by severing the link between justice and cooperation. Thus I will be stating the position somewhat baldly, without providing much in the way of argument in support of it, much less tying up all the loose ends.

My earlier presentation of “minimally controversial contractualism” followed the conventions of the genre, in that it treated the contract thought-experiment as though it were intended to provide foundations for what Annette Baier has referred to as a *normative theory*, viz. “a system of moral principles in which the less general are derived from the more general.”³¹ This is the standard contractualist view: the constraints that must be satisfied in order to achieve agreement are used as the basis for a derivation of one or more extremely general principles, such as efficiency or equality, which then serve as “supernorms,” from which more specific norms, such as “don’t lie,” or “don’t steal,” can be derived. Normative authority flows down, as it were, from the more to the less general (the same way that it does in a Kantian or a utilitarian view).

An alternative way of interpreting these very general principles is to treat them as expressive vocabulary (in Robert Brandom's sense of the term³²) that we introduce in order to talk about broader

31 Annette Baier, *Postures of the Mind* (Minneapolis: University of Minnesota Press, 1985), p. 232.

32 Robert Brandom, *Making It Explicit* (Cambridge, MA: Harvard University Press, 1994), pp. 105-107. “Expressive” is a

patterns in our practices of normative inference. According to this view, primary normative authority rests with the low-level moral norms, which form part of a complex artifact that is reproduced through cultural inheritance. Thus we learn, from our parents, teachers and peers, a set of specific rules to govern our conduct in everyday life, which include a wide range of techniques for managing conflict and creating habituated patterns of prosocial behavior. This is the standard repertoire of rules that any parent is familiar with: not to hit people and grab things, how to form a queue and wait one's turn, techniques for allocating goods both divisible (“you cut I choose”) and indivisible (“eeny-meeny-miny-mo”), deference to legitimate authority, suppression of our tendency to enjoy cruelty, and so on. In principle, there need not be any commonalities between the way that one type of situation is handled and the way that we approach some other. In other words, a culture might have (and many do have) a very specific way of dealing with one area of social life (e.g. marital obligation) and another quite different way of dealing with some other area (e.g. social labour), so that if one were to ask in general terms “what we owe to each other,” the answer would simply be “it depends.” As both formal models of cultural transmission and generations of ethnographers have shown, cultural inheritance is able to sustain almost anything as a normatively enforced pattern of behavior, and consistency across domains tends not to be an important feature (or at least not upon surface inspection).³³

Nevertheless, since culture forms a system of “descent with modification,” it exhibits evolutionary dynamics that are in several key respects comparable to those that prevail in the biological realm.³⁴ Each social norm must compete with other variants that inevitably crop up, both from internal deviance and dissent, as well as from contact (and often conflict) with other cultural groups. The structure of this competition, however, is not neutral with respect to all variants. Our innate psychological dispositions, for instance, although seldom determinative, certainly make some patterns statistically more likely to be reproduced than others. (For example, while compulsory incest and compulsory incest-avoidance have both been normatively enforced at different times in different societies, incest-avoidance has been far more common as a norm. Similarly, while there have been and

somewhat unfortunate choice of terms, for the purposes of reflecting on moral vocabulary, because “expressive” in Brandom's sense of the term has nothing to do with the “expressivist” tradition in metaethics. In Brandom's sense of the term, first-order, thick moral concepts are not expressive. It is only the vocabulary that gets introduced in order to talk *about* these first-order judgements that is expressive (so, for example, words like “ought,” which allow us to transform imperatives into assertions, and thereby embed them in conditionals).

33 Robert Boyd and Peter J. Richerson, “Punishment Allows the Evolution of Cooperation (Or Anything Else) in Sizeable Groups,” in *The Origin and Evolution of Culture* (Oxford: Oxford University Press, 2005): 166-188; see also Donald Brown, *Human Universals* (New York: McGraw Hill, 1991).

34 Of course, this means that it is in many other respects non-comparable. See Peter J. Richerson and Robert Boyd, *Not by Genes Alone* (Chicago: University of Chicago Press, 1995), p. 69. The most important difference is what they refer to as “guided variation.”

are societies that practice polyandrous marriage, polygyny has been by far the more common norm.) The fact that we are inclined to find certain actions easier, or more gratifying, or more repulsive, shows up as a *bias* (in the non-pejorative sense of the term) in the cultural inheritance system.

While our innate psychology provides a set of *content biases*³⁵, cultural evolution is also subject to a set of *pragmatic biases*, which arise out of the structure of social interaction. I would argue that the pragmatic considerations that speak in favor of the two principles of justice outlined above, efficiency and equality, also favor norms exhibiting those properties in the process of cultural evolution. Why? Because these norms favor arrangements that attract fewer complaints, and thus less motivated dissent, than any of their near rivals. Social norms, despite being enforced, still require – as a matter of sociological fact – high levels of voluntary compliance. In other words, they must be able to attract agreement – not necessarily consensus, but at least high levels of agreement. This is because the punishment system itself is stable only to the extent that it is normatively enforced (particularly so when it is decentralized and informal). Thus a certain willingness to play along in good faith is essential to the stability of normative systems. The more they attract objections, the less stable they will be, and the more they will tend to be replaced by systems that attract fewer objections.

This basic framework for understanding the principles of justice has been proposed by Jürgen Habermas, although unfortunately without much uptake. Partly this is because Habermas moves beyond the pragmatics of social interaction to make a series of more controversial claims about the way that structural features of linguistic practice bias cultural transmission.³⁶ There remain, however, important similarities between his account and the more minimal one sketched out here. First of all, both views hold that the principles of justice – in this case efficiency and equality – have no intrinsic normative authority, they merely articulate the end result of a process of cultural evolution. What is being posited is nothing more than a *bias*, the consequences of which are only felt in the fullness of time, and only when not overridden by other forces. But because the pragmatic features of interaction that favor efficient, egalitarian norms obtain in multiple domains of social interaction, all sorts of different practices will tend to evolve in the same direction, or in such a way that they exhibit certain shared structural features. The result is that we are able to make some very robust generalizations about

35 Shaun Nichols, for instance, has argued that moral sentimentalism is, in effect, an illusion produced by the operation of such biases in the reproduction of social norms – such that norms with greater “affective resonance” have a better chance of reproducing. See his *Sentimental Rules* (New York: Oxford, 2004).

36 Habermas claims that: “when it becomes linguistically channeled, social reproduction is subject to certain structural constraints; and.. by reference to these we can – not causally explain, certainly, but – render reconstructively comprehensible, in their inner logic, the... structural transformation of worldviews, the universalization of law and morality, and the growing individuation of socialized subjects.” *The Theory of Communicative Action, Vol. 2*, Trans. Thomas McCarthy (Boston: Beacon Press, 1981), pp. 86-87

“what we owe to each other.” But this is not because our more specific obligations are derived from the more abstract principles that we use to articulate these obligations; it is because the abstract principles were introduced as a way of talking about (in particular, generalizing about) the more specific obligations.³⁷ (Among other things, this framework is also able to explain why, within a culture such as our own, there can be high levels of convergence around low-level moral judgments, combined with deep and persistent disagreement over abstract principles.³⁸)

Of course, once we have developed the expressive vocabulary (having achieved what Brandom calls “semantic self-consciousness”³⁹) we are able to engage in an explicit attempt to direct the evolution of norms in the direction of increased efficiency and equality. For instance, we are able to make use of the principles of justice in a self-consciously “political” fashion, in cases where we recognize the need to minimize disagreement. And when designing and implementing *new* systems of cooperation, we can do so in a way that reflects an explicit concern for equality and efficiency. This amplifies the force of what evolutionary theorists refer to as “guided variation,” further biasing cultural evolution in the direction of contractalist norms, and further enhancing the generality and authority of those norms. This autocatalytic process, which was triggered in our society by the Enlightenment and the emergence of liberal political orders⁴⁰, has provided the central dynamic of moral transformation in our society, and is what accounts for the extreme instability of moral ideas in past century.⁴¹ Thus

37 For further discussion, see Joseph Heath, *Following the Rules* (New York: Oxford University Press, 2008), pp. 272-73

38 Albert R. Jonsen and Stephen Toulmin, in *The Abuse of Casuistry* (Berkeley: University of California Press, 1988), explain their return to casuistic methods through a practical illustration of this phenomenon. The example arose from the participation by one author in a commission struck by the United States government, in order to provide guidance on various bioethical questions. Commissioners were intentionally chosen with an eye toward diversity in several dimensions: “men and women; blacks and whites; Catholics, Protestants, Jews and atheists; medical scientists and behavioral psychologists; philosophers; lawyers; theologians; and public interest representatives”(p. 17). Expecting a high level of disagreement, what he found instead was a fair degree of convergence on practical questions. “The locus of certitude in the commissioners' discussions did not lie in an agreed set of intrinsically convincing general rules or principles, as they shared no commitment to any such body of agreed principles. Rather, it lay in a shared perception of what was specifically at stake in particular kinds of situations. Their practical certitude about specific types of cases lent to the commissioner's collective recommendations a kind of conviction that could never have been derived from the supposed theoretical certainty of the principles to which individual commissioners appealed in their personal accounts. In theory their particular concrete value judgments should have been strengthened by being “validly deduced” from universal ethical principles. In practice the general truth and relevance of those universal principles turned out to be less certain than the soundness of the particular judgments for which they supposedly provided a ‘deductive foundation’.” (p. 19)

39 Brandom, *Making It Explicit*, p. 384.

40 What I take to be fundamental here is Pierre Manent's idea that liberal societies are the first societies to be organized according to an idea of how a society should be organized. See his *An Intellectual History of Liberalism*, Trans Rebecca Balinski (Princeton: Princeton University Press, 1995), p. xv-xviii.

41 The result has been as Michele Moody-Adams describes: “Much of the moral language that helps shape the economic, social, and political dimensions of the contemporary world is a product of distinctively philosophical efforts to articulate interpretations of the structure of moral experience.”Michele Moody-Adams, *Fieldwork in Familiar Places* (Cambridge, MA: Harvard University Press, 1997), p. 194.

general normative principles, in this view, are not extra gears, merely by virtue of being expressive. But they are considerably less central to the mechanism of moral judgment than many philosophers have taken them to be.

IV

So how does this way of looking at things help to solve the puzzle? It does so by suggesting that the egalitarian intuitions deployed by the average person in our society are neither “built up” from a set of principles governing individual interactions, nor are they “inferred down” from a conception of how society as a whole should be ordered. They are instead a product of simultaneous, convergent cultural evolution in different domains of interaction, catalyzed by the reflexive use of interpretive vocabulary in practices of social criticism and reform. This is why our egalitarian intuitions across different domains only sometimes add up in a way that is consistent with egalitarianism at the level of society as a whole – it is because they are not derived from a commitment to an abstract principle of equality.

Adopting this perspective helps to explain some of the peculiar wrinkles in the way that we apply egalitarian ideas. Take, for example, what has come to be known as the “Titanic puzzle.” It arises from a rather casual remark in Thomas Schelling's *Choice and Consequence*, in which he suggested that the Titanic had an inadequate number of lifeboats because passengers in 3rd class (or “steerage”) were expected to “go down with the ship,” and that this was somehow part of the conditions of carriage associated with the less expensive tickets.⁴² The puzzle is then as follows: assuming that we find it unacceptable for passengers on the same boat to have differential access to lifeboats, on the grounds that some did and some did not pay for this safety feature, how then can we accept an arrangement under which passengers on *different* boats, having paid different prices for carriage, have access to different levels of safety?⁴³ (After all, different ships provide different levels of safety, in the same way that different automobiles do.)

The standard micro and macrocontractualist frameworks seem unable to capture what is troublesome about this case. From a micro perspective, there would seem to be nothing wrong at all with passengers on the same ship having differential access to lifeboats – indeed, insisting that there be enough lifeboats for everyone is guaranteed to create a deadweight loss, since it will raise the price of

⁴² Thomas Schelling *Choice and Consequence* (Cambridge, MA: Harvard University Press, 1984), p. 115.

⁴³ See Debra Satz, *Why Some Things Should Not Be For Sale* (Oxford: Oxford University Press, 2010), p. 88. I hesitate to use this example, because it risks perpetrating an urban myth, since the account of conditions on the Titanic is entirely fictitious (indeed, the suggestion that there was a policy of denying 3rd class passengers access to the lifeboats was vehemently denied by White Star Lines).

tickets, thereby making them unaffordable to a thin slice of consumers who would have been willing to pay just slightly less for carriage, and who could have been squeezed onto the ship. If a group of passengers enters into a transaction with White Star Lines that is mutually beneficial, and the terms of that transaction are internally just, how can it be relevant that some other group of passengers enters into a different transaction with White Star Lines, which is also internally just, but contains a different set of terms? There is no presumption of equality between the passengers with respect to lifeboat access, because the passengers are not cooperating *with one another* in order to secure the provision of this good.

The macrocontractualist is in a better position to criticize the “steerage goes down with the ship” arrangement, because he can say that, from the standpoint of society as a whole, it is a violation of equality for some people to be exposed to mortal dangers that others are able to protect themselves against. But then she is unable to explain why we are untroubled by the fact that *different* ships have different safety standards, and that passengers might choose one ship over another because these differences resulted in a lower price of transport. What is it about being on the *same* ship that somehow makes it troublesome for access to lifeboats to be distributed in accordance with ticket price?

Rather than searching for general principles from which this particular constraint can be derived, there is much to be gained simply by noting that shared transportation is a particular type of cooperative enterprise, which is subject to a distinct set of norms that have evolved and adapted over time. Typically these norms require greater forbearance than is expected in everyday interaction – willingness to tolerate greater encroachment of one's personal space, deference to the authority of the captain or driver, restrictions on one's ability to engage in activities that might jeopardize the safety of others, or slow down passage, and also a set of procedures for dealing with emergencies. It is the latter set of norms, I would suggest, that generates the “puzzle” in the Titanic scenario – the view that it is impermissible for passengers on the same boat to have differential access to lifeboats, even though there is no general social requirement that safety on different ships be equalized. Emergencies typically evoke a higher level of social solidarity than everyday interactions, and so the norms governing them are often more egalitarian. (This is also true of face-to-face interactions, such as G. A. Cohen's “camping trip.”⁴⁴) As a matter of historical record, the norm that actually governed the evacuation of the Titanic was “women and children first” – to the point where men were barred entirely from entering lifeboats on one side of the ship. (Indeed, Schelling's claim that passengers in steerage were expected to

44 G. A. Cohen, *Why Not Socialism?* (Princeton: Princeton University Press, 2009).

go down with the ship is simply false. Survival rates among women even in steerage were much higher than that of men, including those in first class. The lower survival rate of passengers in steerage can be almost entirely explained by the lower percentage of women traveling third class, along with the physical positioning of the lifeboats on the upper decks of the ship.) This is, one might note, not exactly an egalitarian norm – it discriminates on the basis of gender and age. It is just that within the different groups (men, women, children), it does not permit further discrimination (based on, say, seating class), but rather applies a queuing norm of “first come, first served.”⁴⁵

The best way of thinking about this example, I would suggest, is to regard “travel by ship” as a particular sort of cooperative practice, governed by a distinctive set of norms (i.e. “naval tradition”). Differential access to lifeboats, based on carriage class, violates these norms, in the same way that trying to buy your way into line at a movie theatre violates the norms governing the practice of queuing. When one is dealing with different ships, on the other hand, the same norms do not apply. (There are of course *different* norms that apply between ships, such as the obligation to divert course in order to effect a rescue. These are part of a system of generalized reciprocity that has a cooperative structure, but the overall objectives are different.)

The idea that particular sets of norms are tied to particular systems of cooperation remains important, in part because it explains why egalitarian principles tend to be “local” in their application.. The cultural-evolutionary model proposed above suggests that distributive obligations are likely to be limited in scope to those who are directly involved in the cooperative scheme, simply because they are the ones whose voluntary compliance needs to be secured, and therefore, the ones whose complaints need to be addressed in order to secure reproduction of the norm. Thus it is hardly surprising to find “special-purpose” norms that are adapted specifically to regulate the behavior of passengers on the *same* ship, since these are precisely the people who need to cooperate with one another in order to get various things done. When these local norms are applied, the result will be consistent with the way that microcontractualists have tried to model the application of principles of justice. Of course, there is no way to draw a crisp line to demarcate the scope of a cooperative enterprise. This is, however, not a specifically *philosophical* problem, because individuals’ own ability to manage their daily affairs depends upon an ability to classify interactions using the appropriate “cultural template,” and to apply the relevant norms. The scope of their obligations is ultimately defined by these norms. The specifically philosophical project of offering a rationale for these norms, using abstract concepts such as

45 On this, see Elster, *Local Justice*, pp. 73-74.

“cooperation,” and “equality,” should not be seen as an attempt to unearth the reasons why people have the intuitions that they have about specific cases. On the contrary, the philosophical articulation and refinement of these norms primarily serve to enable a more “clairvoyant” continuation, or critique, of the practice.

The reason that we seek to apply egalitarian norms at a more abstract level – and so worry about things like “the distribution of wealth” or “gender inequality” in society as a whole – is twofold. First, the development of the modern nation-state has generated systems of cooperation that genuinely do encompass all members of society. Once a uniform system of property rights is put into place – a system chosen from an enormous menu of options, each with different implications for the production and the distribution of goods – then the consequences that this system has for inequality in society as a whole becomes a legitimate object of concern and critique. The same sort of consequences flow from the development of universal childhood education, conscription, comprehensive health insurance, increased legal regulation of family relations, and so on. The second reason is that we have been and continue to be engaged in the project of constructing institutions that will enable us to cooperate on a larger and larger scale. This is being done self-consciously, as it were, explicitly using the principles of efficiency and equality in our processes of institutional design, precisely out of a recognition that we lack the sort of shared cultural resources that could bring about consensus around some thicker set of norms. Thus we articulate our aspirations, in terms of collective actions problems that we would like to see solved and systems of cooperation that we would like to see institutionalized, in the contractual language of efficiency and equality. This exercise is going to look like macrocontractualism, in the sense that both the principles and the system of cooperation are going to be conceived of in a very stylized way (so, for example, we may talk about controlling global warming as a “public good,” even though, strictly speaking, not everyone benefits).

As a result of these institutional developments, we wind up with a set of normative commitments that are, if not formally inconsistent, then at very least in tension. We tend to have strong views about how particular interactions should be organized, in order to meet certain standards of justice, but we often become uncomfortable with the aggregate consequences of organizing these interactions in this way, and so cast about for ways to tweak or rearrange things, so that the large-scale outcome is one that satisfies our conceptions of justice at that level. Reconciling these tensions is not primarily a philosophical problem, it is a practical problem – that of finding ways to bring our institutions into line with our ambitions and ideals, so that they form an integrated and consistent

system.

V

Philosophical debates about the requirements of justice have been drifting, over the course of the past few decades, toward increasingly abstract concerns about “equality” at the level of society as a whole, or the entire human race. Those who have resisted this tendency have attracted a degree of opprobrium, based perhaps on the suspicion that, when push comes to shove, the real basis for their opposition to this expansive conception of equality is that they are rich Westerners who don't want to share their stuff. Examining the case more charitably, however, it quickly becomes apparent that there are a variety of motives. Perhaps the most common reason for concern is that it makes the theory of justice utopian in the pejorative sense of the term (and therefore useless when it comes to addressing any real-world political questions). Elizabeth Anderson has suggested, along these lines, that “in focusing on correcting a supposed cosmic injustice, recent egalitarian writing has lost sight of the distinctively political aims of egalitarianism.”⁴⁶

There are, however, somewhat narrower, more philosophical reasons for concern. The major one, which I have focused on here, is that by applying the principle at this abstract level, one loses sight of the role that equality – or more diffuse conceptions of fairness – play in mediating interpersonal relations and institutional decision-making. University professors, for instance, care very much what sort of salary the person down the hall is drawing, and tend not to evaluate the fairness of that arrangement in terms of its contribution to global equality. The principal of a school, faced with excess demand for enrollment, is concerned to implement impartial admission procedures, but feels no need to calculate what impact that this will have on inequality in society as a whole. A group of entrepreneurs, going into business together, settle on a division of ownership shares based on a conception of fairness, typically one that ties contribution to reward *among the partners*. The examples can easily be multiplied.

The most natural account of what people are doing when they apply these sorts of norms, and in particular, why they gravitate toward egalitarian norms, is that these norms are conflict-minimizing – as, indeed, anyone who has participated in this sort of decision-making can attest. (The best way to discover the attractions of egalitarian norms is to experience the consequences of *failing* to apply egalitarian norms.) This is the intuition underlying the contractualist idea that the principle of equality

46 Elizabeth Anderson, “What is the Point of Equality?” *Ethics*, 109 (1999): 287-337 at 288.

(and, more obviously, the principle of efficiency) might be valued for its ability to bring about agreement in cooperative enterprises. Developing this plausible-sounding idea into a fully specified theory, however, has proven difficult. If one treats contractualism as having the structure of a “normative theory,” with a single set of very abstract principles which then get applied in a way that generates more specific normative constraints, then one quickly winds up caught up in the puzzle described in the first two sections: there is no way to establish a consistent micro-macro link. It can be avoided, I have suggested, by instead adopting a cultural-evolutionary framework, and viewing micro-macro consistency as a *social project*, rather than as a logical requirement of normative theories.