CONSERVATIVE CONTRACTARIANISM

By

Terrence C. Watson

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Arts
in
Philosophy

Waterloo, Ontario, Canada, 2004

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

CONSERVATIVE CONTRACTARIANISM

ABSTRACT

Moral contractarianism, as demonstrated in the work of David Gauthier, is an attempt to derive moral principles from the non-moral premises of rational choice. However, this contractarian enterprise runs aground because it is unable to show that agents would commit to norms in a fairly realistic world where knowledge is limited in space and time, where random shocks are likely, and where agents can be arbitrarily differentiated from one another. In a world like this, agents will find that the most "rational" strategy is to behave "non-rationally," imitating the behavior of others in their vicinity and preserving a limited sort of ignorance.

ACKNOWLEDGEMENTS

First: Jan Narveson, for teaching me all philosophy should be read as if written just last week by someone down the hall.

Second: Bill Abbott and Brian Orend; the former for introducing me to game theory in a fairly rigorous way, and the latter for giving some of the best lectures I have ever attended.

Third: Peter Jaworski and Amanda Chalmers; in different ways, both were instrumental in the formation of the better ideas in this thesis. The bad ideas, alas, are mine and mine alone.

Fourth: Debbie Dietrich and Linda Daniel; thank you both for being so patient with me!

TABLE OF CONTENTS

TABLES & ILLUSTRATIONS

1. Introduction: Lofty Goals

While there is no single, unified contractarian tradition, contractarians address the same question: why should we be moral? Before we present their replies, it would be productive to attempt to answer it ourselves. Are you already a good person? If so, then perhaps you need no further argument to maintain your standard of good behavior, and this present inquiry will be pointless to you. If you *are* good, and are fortunate enough to be surrounded by others who are mostly as good as you are, then contractarianism has very little to offer you.

But let us suppose you are not a good person. Indeed, let us suppose you are evil, not for the sake of being evil, but because you have learned that a few thefts or lies – or a few murders? – make it easier to achieve your other ends. Mostly, you perform your little evils only in those cases where you are unlikely to be caught, when no one is looking, or when someone else will be blamed. No, you are not a good person, but your consistent pursuit of your interests and your ability to avoid getting caught indicate that you are a reasonable one.

The question: why should you be moral? Or, to put it another way, what convincing reason could I give you to become a good person? It is important to me that you become a good person for at least two reasons: As long as you are lying, stealing, or killing and *not being apprehended*, you are a threat to me, both directly, in case I get in your way, and indirectly, if you lead others to an evil lifestyle. In this case, the contractarian argument is a *defensive* posture; it is beneficial to me, in the most straightforward sense, for you to become a good person. This would eliminate both the costs of enforcement and the most direct costs associated with becoming your prey.

Second, it is important to me that you become a good person because *that is what my idea of the good demands*. Just as the Form of the Good compelled the philosopher to return to the cave and share his wisdom with everyone else, it seems an inevitable requirement of morality that I convince you to accept my idea of morality. By seeking reasons to be moral that should be acceptable to the widest group possible, contractarianism attempts to fulfill this obligation. At the same time, a universally compelling argument is the best protection against evil a theorist could hope to find.

These are the basic characteristics of a contractarian argument, but here the similarities tend to end. Boucher and Kelly call the sort of argument we have just described "moral contractarianism", suggesting it is an attempt to "ground moral principles in the creative self-interest of individuals who adopt constraints on their behavior in order to maximize benefits." Robert Sugden [1990] claims that, "Contractarians seek to derive principles of morality by analyzing the problem that would be faced by rational individuals in a state of nature." As David Gauthier [1986, p. 17] argues, a contractarian theory "enables us to demonstrate the rationality of impartial constraints on the pursuit of individual interest to persons who take no interest in others interests."

Each of these descriptions generally fits our characterization of the contractarian argument, but each provides ample scope for differences among theorists. For example, we might further inquire into the motivations "creative self-interest" presupposes or allows into the contractarian analysis. We might also differ in our conception of the state of nature and why it is so problematic to rational individuals. Finally, we may question the nature of the impartiality of the constraints a contractarian theory derives. Still, even

with these differences, we can present a preliminary sketch of the basic contractarian argument: "Given certain motivations, and placed in an initially inefficient situation, subject to certain limitations to ensure the impartiality of the contract, agents would be rational to agree to certain constraints on their behavior." According to the argument, under those conditions – and possibly only under those conditions – we could describe the agreed upon constraints as *moral* principles. This is the general argument that this current treatise intends to criticize.

Of all modern contractarian theories, David Gauthier's most clearly fits the above description. It is an example of moral contractarianism *par excellence*, an attempt to defy Hume's Law and derive morality from premises that are completely without moral content. "Morality," he argues [1986, p. 4], "can be generated as a rational constraint from the non-moral premises of rational choice." If this claim is correct, then when placed in an appropriate state of nature setting, even our rational evildoer would agree to constrain his behavior in certain ways. Being a concrete example of our model argument, Gauthier's contractarianism will be the focus of the criticism in the next section, where we will demonstrate the inadequacies of the general contractarian enterprise.

The next section will be organized into four parts: first, we will assess Gauthier's rational agents. Second, we will examine his state of nature. Third, we will take a look at the limitation he places on the initial position in order to ensure the impartiality of the ensuing agreement, the so-called Lockean Proviso. Finally, we will review the agreement itself, especially its hypothetical nature, in an attempt to understand the bearing it might have on morality. At each stage, we will suggest ways in which Gauthier's contractarian argument might be improved. When the ground is cleared, we

will expand this part of the critique and offer a more complete, comprehensive alternative to Gauthier's vision that still fits the spirit of the contractarian model.

2. Rational Agents

In her essay on Gauthier's contractarianism, Margaret Moore [1994, p. 211] claims that his model agent is a "rational, self-interested (non-tuist) utility-maximizer." We should take this to imply that Gauthier's argument is intended to convince people who may or may not care about the interests of others that rationality obliges them to accept certain (moral) constraints on their behavior. As Moore argues, "Obviously different conceptions of what is essential to the person will yield different sets of principles or rules which would be acceptable to persons so described. It is also crucial that the parties to the contract are people with whom we can identify." Everyone has self-interest, but not everyone has an interest in the welfare of other people.

In this way, Gauthier's utility maximizer simply has goals and acts in a manner that is most likely to achieve those goals. Sometimes, the most efficient way to achieve a goal may be to steal, maim, or kill other agents who are similarly rational. The effect of such straightforwardly rational behavior is to create a situation commonly described as a Prisoner's Dilemma: it's good for me if no one gets maimed or killed, but even better for me if you get maimed or killed and I do not, and likewise for you. This leads to a situation that is worse for both of us than the one where no one gets killed or maimed at all. This is the state of nature, which we will examine in detail in the next section.

Gauthier's agents are not just rational but *equally* rational. First, they all act in the ways most likely to achieve their ends. But also [Moore 1994], "Being equally rational – with no psychological strengths and weaknesses – they are all able to detect dispositions with roughly the same degree of accuracy. And, being equally rational, agents will presumably reason identically." So agents not only act to achieve their ends,

but given the same ends and similar situations, they will act in the same way. Moreover, and most controversial for many commentators, the agents are capable of determining with a high degree of accuracy the behavioral dispositions of those with whom they interact. This is what Gauthier means when he supposes that dispositions will be "transparent" or "translucent."

This section is concerned not so much with Gauthier's model of rational choice through utility maximization, but more with the compatibility of that model with the notion of a "moral disposition" he wants to promote. Moore [1994, p. 216] points out that "Gauthier is using the assumption [of equal rationality] to mask important differences between people, differences in their talents and abilities and preferences, those things on which rational agents usually *base* their decision about what it is rational for them to do." By this, she seems to mean that two people could be equally *rational* (i.e. both could seek to maximize their utility) even while one is a better deceiver than the other, and that in their dealings with each other these two equally rational people could make different decisions.

But this assumes that we can make sense of the idea of what it would mean for a rational agent to have a disposition in the first place; for if dispositions are incoherent, then rational agents would have nothing to deceive each other about (as far as Gauthier's theory goes), even if deception was possible for them. And if deception is impossible, then we need to know what a disposition consists of before we could know exactly what the transparency assumption would imply. If the idea of a rational agent having a disposition makes no sense then the problems with Gauthier's theory run very deep

indeed. As will be shown, these problems have implications for the very core of the contractarian endeavor.

According to Gauthier [1986, p. 167], "A constrained maximizer has a conditional disposition to base her actions on a joint strategy, without considering whether some individual strategy would yield her greater expected utility." A constrained maximizer, therefore, has a disposition to "co-operate in ways that, if followed by all, would yield outcomes that, if followed by all, would yield outcomes that she would find beneficial and not unfair, and she does co-operate should she expect an actual practice or activity to be beneficial."

But what does it mean to possess a disposition to cooperate? It is essential for Gauthier [1988, p. 177] that moral constraints are internal, "operating through the will, or decision making, of the agent," and that they operate in a way "that satisfies some standard of impartiality among persons." But there is more. For the internalization of moral constraints cannot just be a function of "particular psychological phenomena, however benevolent and humane their effect, and however universally they may be found." [1986, p. 103.] Supposedly, then, we should be able to make sense of constrained maximization *solely* within the framework of rational choice theory. Yet, as Ken Binmore [1994] points out, if the mechanics of the theory are to be preserved, then constrained maximization *cannot* work the way Gauthier wants it to work.

Binmore's point is that the payoffs in any game, like the single shot Prisoner's Dilemma game Gauthier analyzes, must be interpreted according to revealed preference theory. This means that the payoffs are *not* measurements of pleasure or of monetary reward or anything like that. As Binmore [1994, p. 105] emphasizes, "[Economists]

regard it as a *fallacy* to argue that a person prefers one thing to another *because* the utility of the first exceeds the utility of the second…A player's payoffs in a game are *deduced* from his preferences over its possible outcomes." Preferences are themselves revealed through observed behavior. Thus, in the game depicted below, the *reason* (defect, cooperate) has a higher payoff than (cooperate, cooperate) is because when the player has to choose between the two he chooses the first and not the second.

	Cooperate	Defect
Cooperate	2	0
Defect	3	1

Figure 1: Prisoner's Dilemma With Cardinal Payoffs

For identical reasons, (defect, defect) has a higher payoff than (cooperate, defect.) Thus, the player defects when he knows his opponent is going to defect and also defects when he knows his opponent is going to cooperate. But this makes the very notion of what it would mean to have a disposition troublesome. If being disposed to constrained maximization means that you would cooperate if you believed your opponent would cooperate, then the payoffs in the game – and the game itself – would have to change.

However, Gauthier is adamant that the structure of the Prisoner's Dilemma be retained, but within that structure it is unclear what having a disposition could possibly mean. On the simplest level, suppose I am playing the Prisoner's Dilemma with someone who is "disposed" to cooperate no matter what he or she believes about your own "disposition." Again, if I am rational and believe my opponent will cooperate, I will defect; that is what the payoffs in the matrix *mean.* If I had cultivated within myself a disposition to be nicer than that, then the payoffs will be different, and we would be

playing a different game. Actually, for my opponent to possess this disposition makes the game different, supposing he is rational.[1]

Now suppose you are playing with a constrained maximizer. "A constrained maximizer," Gauthier [1986, p. 170] explains, "chooses to co-operate if, given her estimate of whether or not her partner will choose to co-operate, her own expected utility is greater than the utility she would expect from the non-co-operative outcome." For the purposes of this example, we must presume that I have not chosen my own disposition yet; I will choose my disposition (or not) and *then* play the game. Once I have chosen a disposition, I cannot change it, at least until after the game is played. This *seems* to mean that I am *committed* to my disposition, or at least to the course of action it demands of me. Our argument here is that it is inconsistent with rational choice theory to presume that I am able to commit myself in this way.

So now I have disposed myself to cooperate with likeminded constrained maximizers and, because of Gauthier's transparency assumption, my opponent is aware of my disposition. My opponent will now cooperate – and, most problematic of all – now I have reason to *believe* she will cooperate. I believe that, "She believes I am a constrained maximizer, so she will decide to cooperate with me." Because of the argument already given, I will now defect. Gauthier seems to want to say that constrained maximizers would simply be *blind* to at least one of the cells in the payoff matrix of the Prisoner's Dilemma, but clearly, if you eliminate that box from a player's point of view, he is no longer in a Prisoner's Dilemma. Instead, the players would be playing a game like the one depicted below.

[1] Perhaps the other prisoner is Jesus Christ or some other saintly figure! Maybe it is impossible for saints to play the Prisoner's Dilemma at all.

	Cooperate	Defect
Cooperate	1,1	--
Defect	--	2,2

Figure 2: Constrained Maximizer Coordination – A Game That Is Not The PD

It will doubtlessly be argued that, for (cooperate, defect) and (defect, cooperate) to remain possibilities for me, I must not have *really* committed myself to constrained maximization. But it is completely unclear what an *actual* commitment would consist of in this case, except perhaps if it entailed a sort of conditioning that changed the game's payoffs.

It is attractive to assume that agents in the state of nature ought to be able to commit themselves to any course of action they like. For example, suppose that anyone at any time can commit himself to a particular strategy, such that everyone else can infallibly recognize the commitment that has been made. Thus, if in the Hobbesian state of nature I promise that I will play like a dove until you play like a hawk, you can believe my promise without fear of deception. Perhaps, to satisfy Gauthier's transparency assumption, whenever I commit myself to a strategy, a note appears above my head that everyone else can read. People who haven't made any particular commitment don't have a note. Of course, on this common sense idea of the state of nature, it seems rational to be suspicious of those who don't have notes; indeed, one might almost make a commitment to default to playing hawk with those who have not made any commitments themselves.

So far, not much has been said about the game that people are playing in the state of nature, but this is because it needs to be made clear that the capacity for commitment can be defined *extraneously* to the game theoretic model. This simplifies the task for the contractarian: a note that explains that one will play dove with those who have

committed to playing dove and hawk with those who have committed to playing hawk (or, more likely, have made no commitments at all) is immensely valuable, no matter what game has been chosen to represent the state of nature. Moreover, it necessarily serves as a constraint on behavior, preventing a player from taking advantage of others who will be using the dove strategy (whether as a default or because they believe the commitment sign.) The long-term benefit of having others know you've made a commitment to be peaceable outweighs any short-term gain from lying, stealing, or killing.

As Gauthier [1986, p. 183] argues, "A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition…The essential point in our argument is that one's disposition to choose affects the situations in which one may expect to find oneself." Our telltale notes seem to satisfy at least some of the requirements of a disposition. And it is true that costless commitment of the type we have described here would drastically influence the kinds of situations agents would find themselves in. With clear distinctions between good players and nasty ones, the inefficiency of *any* state of nature diminishes. Trust, if you want to call it that, becomes a virtually costless commodity, so while some resources may still be spent on fighting, they will be spent on fighting the right people (hawks and the like) and not other people who are willing to cooperate. Devoted hawks will find themselves meeting only other hawks (people who have made a commitment to play hawk with other hawks.) As Brian Skyrms [1996, p. 21] notes, "If there is some tendency, for whatever reason, for like-minded individuals to interact with each other then the prospects for the evolution of justice are improved." By

allowing commitments to be made, we are drastically increasing the probability that the dove strategy will only be played against other doves and that the hawk strategy will only be played against other hawks.

If we suppose that the state of nature is a Prisoner's Dilemma game played repeatedly, but with players paired at random so that no one can remember who he has played with before, then it is easy to deduce what will happen to that state of nature.[2] Those who have made a commitment to play dove with those who are committed to playing dove and hawk otherwise will do very well. They will cooperate with doves and people who have similar dispositions to their own, reaping the benefits of cooperation, but defect against hawks or those who have made no commitment at all. Indeed, it could be argued, as David Gauthier should, that the rational thing to do is to make a commitment of this type. Hence, starting from a state of nature, it is rational for individuals to accept constraints on their behavior – contractarianism succeeds!

This is what I call "social contracting the easy way." But the contention here is that it is a mistake to assume that players can make binding commitments *independently* of the game being played; hence, the argument, while easy to make, is fundamentally flawed. Along with Ken Binmore [1994, p. 162], we want to say that, "Commitment assumptions…should be built into the rules of the game when the game is constructed." The point here seems to be that commitment assumptions come in at the modeling stage, when the structure of the game is being defined. "[The game theorist] does not allow himself commitment assumptions when *analyzing* a game. To do so is to deny

[2] If commitments of the type being described here are allowed, there is little difference in principle between a random pairing game and an iterated one where strategies like tit for tat can be used. In fact, as will be argued, commitments make cooperation even easier to sustain in a state of nature than the equilibriums formed by reciprocal strategies in the iterated game.

propositions that the game theorists regard as tautological." A rational player cannot make credible, binding commitments in just *any* game, but only in those games that are specifically designed to enable the making of such commitments.

At this point, it might be argued that the only thing that prevents rational players from exhibiting a general capacity to make commitments is revealed preference theory. Our solution, then, would simply be to jettison revealed preference theory. But this is unnecessary and probably unwise; for one, it would still leave us with the very thorny problem of determining how cooperation in a situation that *really* resembled a Prisoner's Dilemma would be possible. A possible response to this problem is to say that such situations never arise, but why should we assume *that?*[3]

As we initially claimed, good people who interact only with good people would have little use for the sort of contractarian argument Gauthier makes. For those who would feel badly about stealing from or killing other people, morality is already rational; such agents have all the reason they need to be moral. However, "Nontuism offers a worse-case scenario. Suppose persons take no interest in the interests of those with whom they interact; nevertheless, they are rationally required to accept constraints on the pursuit of their concerns, and those constraints are based on the interests of their fellows" [Gauthier 1988, p. 213.]

In some sense, then, Gauthier's nontuistic agents serve the same purpose in his theory that Robert Nozick's state of nature served in his. Nozick [1974] selected a pleasant anarchy where people usually respected the rights of others because if he could show that a government would evolve out of implausibly nice initial conditions then he would also have shown convincingly that it would have evolved from circumstances that

[3] Yes, I *do* have a reason to say that, but it has to wait until the second, more constructive part of this work.

were not so nice. Laurence Thomas [1988, p. 154] notes Gauthier's parallel argument in this way: "If it can be shown that it is rational for perfectly selfish people to accept the constraints of morality, then it will follow, *a fortiori*, that it is rational for people capable of affective bonds, and thus less selfish, to do so." If morality can be shown to serve your interests even when your interests do not inherently coincide with morality, then it will support your interests even more when they do coincide with morality to some extent. Thus, having an independent rationale for morality can only increase the likelihood that a naturally good person will behave morally, not diminish it.

For a similar reason, we should avoid taking the easy way to the social contract by making large assumptions about the ability of people to commit themselves to a course of action. As we will show, aside from weakening the outcome of the contractarian enterprise, there is no need to make those kinds of assumptions.[4] We can do as Binmore suggests and build the capacity for commitment *into* the game, rather than presuming that it is simply a property of rational agents regardless of the game being played. This means leaving the Prisoner's Dilemma behind, as well as any other game that takes only a strategic representation. As we will see, commitments often or even always require situations in which one agent chooses his strategy before the other; and the agents who are so committed need not be equal in their positions, nor in their vulnerabilities, for the contractarian enterprise to succeed.

[4] Assumptions like: if people have a conscience, they will cooperate in a Prisoner's Dilemma. Or if they do not want to be known as a defector then they will cooperate. Etc.

3. The State of Nature

Focus on a group of our preternaturally rational agents. Choose two of them at random, placing them near one apple, or one dollar, or one other unit of some resource that both find valuable. They arrive simultaneously. Each would prefer the entire unit and to be left alone by the other; but both would prefer to fight for the resource than to go entirely without it. If they fight, they will both take wounds and at least part of the resource will be destroyed, with the winner taking whatever is left over from the battle. If they agree to split the resource in some way least one of them will be better off than if they fought, possibly both. Demonstrating the dilemma the rational agents find themselves in, this situation has all of the elements of a Hobbesian state of nature. Mutual cooperation would allow them to bargain over the resource and split it efficiently, but it's disastrous to try to bargain if your opponent is only interested in fighting. However, mutual fighting wastes a part of the resource and results in damage. The resources wasted in fighting, relative to the division of the resource that would occur as a result of cooperation, is a measure of the state of nature's inefficiency.

In the above example, the structure of the state of nature is entirely dependent on the outcome of the bargaining game, even if that game is not currently taking place. Whether their strategic conflict is soluble or not depends on what *would* happen if the agents bargained instead of fighting. If the bargaining outcome is expectedly worse for a player than defection, then bargaining will never occur: the agents will go for their guns instead. The bargaining game is what *shapes* the possibilities and problems of the state of nature.

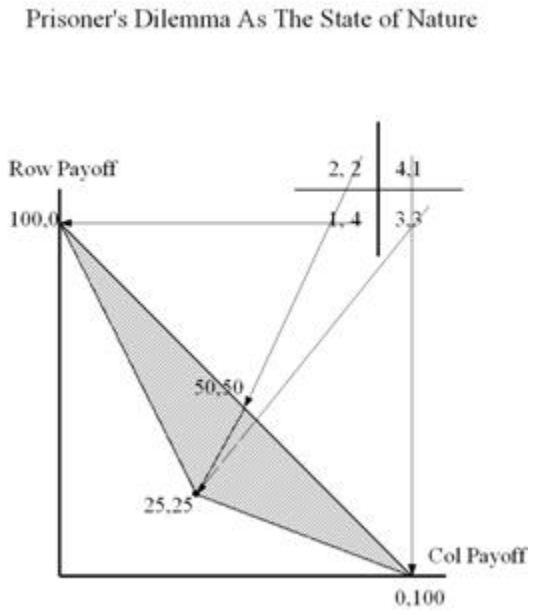

Figure 3: Prisoner's Dilemma As The State of Nature

To illustrate, consider Figure 3, which graphically connects the strategic Prisoner's Dilemma with the underlying bargaining game. The diagram assumes that mutual fighting will automatically destroy half the resource, and that each agent has a 50 percent chance of absconding with the rest after that. This means that if both players fight, the expected utility to both is 25 percent of the resource. If one player fights and the other does not, the one who fights gets all of the resource (100 percent) and the other gets nothing. In the diagram, the line that represents the bargaining game goes from (100,0) to (0,100), representing all possible divisions of the resource if none is lost to fighting. The point at (25,25) represents the expected utility to each agent if both decide to fight.[5] The shaded portion represents areas for improvement, where – at least conceivably – morality might have some work to do, by reducing fighting. The point at

[5] In all the diagrams here, the shape is what is most important. Conceivably, the point representing mutual fighting could have been pushed back almost to zero, and the (hawk, dove) and (dove, hawk) points could have been pulled closer to zero as well. This would distort the shape, but not destroy it completely.

(50,50) was chosen to represent Pareto efficiency, where no further moral improvements are possible. While this is where morality should get us, ideally, this should *not* be taken as an endorsement of egalitarianism; instead, it simply represents the way the resource would be distributed if no fighting took place. The position of the optimal bargaining point will vary with the shape of the bargaining set (curve or line?) and, more importantly, with the amount of symmetry between the players. It only happens to be egalitarian here.[6] In all cases, the distance between the optimal bargaining point and the outcome of mutual fighting measures the inefficiency of the state of nature, but also provides space for that inefficiency to be remedied. This is what we mean when we say that the bargaining game structures the state of nature.

By varying the seriousness of fighting and holding the optimal bargaining point constant, we can change the shape and size of the cooperative opportunities available to the players. For example, suppose that the resource is extremely delicate so that mutual fighting destroys it completely, making the state of nature a game of Chicken (Figure 4.)

[6] What we call the optimal bargaining point could also be called the Nash bargaining solution. The NBS favors the player who has the least to lose if no bargain is struck. It is egalitarian in this case because both players are equally disadvantaged should fighting occur. Also, if we substituted utilities for percentages, the optimal bargaining point would only appear to be egalitarian (50 utils for me need not be the same as 50 utils for you.)

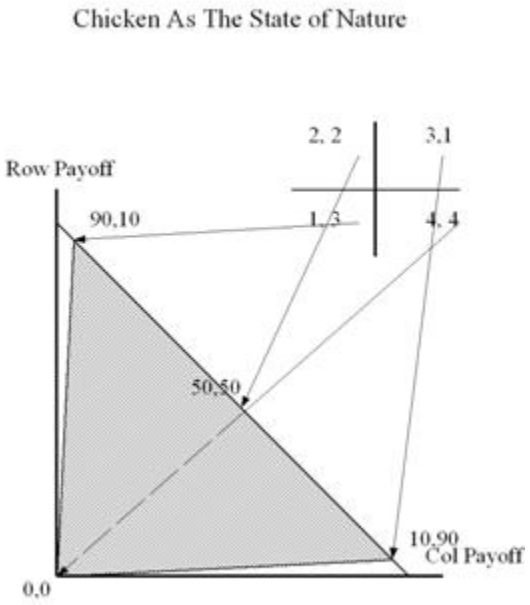

Figure 4: Chicken As The State of Nature

In this case, the space between the conflict outcome and the optimal bargaining point is very large, meaning that there are tremendous cooperative opportunities. Somewhat paradoxically, this isn't necessarily a good thing; it allows one player, basically the first player to make a move, to dictate terms completely to the other player. Even if the optimal bargaining point mandates a fifty-fifty split, if one player commits himself to accepting no less than 99 percent of the resource, the other player has reason to commit himself to accepting the remaining 1 percent.[7]

In the state of nature that is just the opposite from the one above, mutual fighting is only as destructive as in the Prisoner's Dilemma, leaving both players with an expected utility of 25 percent. However, unlike either Prisoner's Dilemma or Chicken, being the

[7] The resemblance to the ultimatum game is not accidental. Chicken makes the bargaining set and the set of feasible bargains (almost) identical. The optimal bargaining point is thus only one of the feasible bargains available to the agents, in contrast to the Stag Hunt game we will examine shortly. In other words, the bargaining area does not make the optimal bargaining point a particularly salient option to the players.

only one to fight is just as destructive as mutual fighting. In this case, perhaps, it's not that taking the resource entirely for oneself is destructive, but that more of it can be had in cooperation with the other player. In any event, fighting to possess the resource nets only 25 percent of it, even if the other player does not fight back. This provides the state of nature with the interesting structure exhibited below. Given the above modifications, the model being used is now the Stag Hunt game.

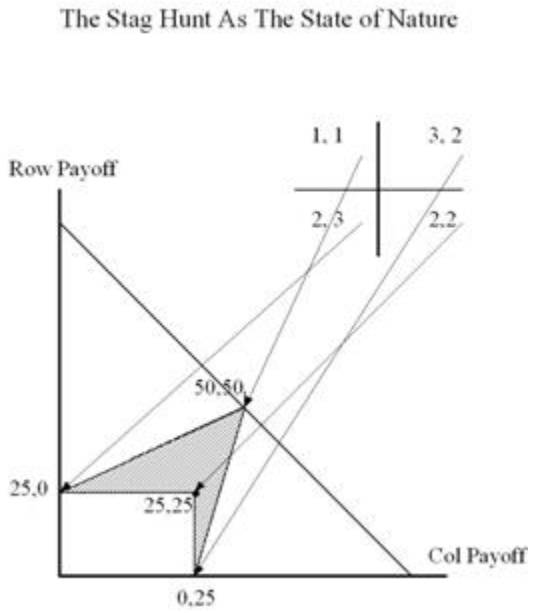

Figure 5: Stag Hunt As The State of Nature

Here the furthermost tip of the bargaining area coincides with the optimal bargaining point. This means that an agent's commitment to the bargaining process instead of fighting should scale with the amount of commitment found in his opponent. Unlike the Prisoner's Dilemma, maximal commitment to bargaining (and not fighting) is advantageous for both players. However, as we have already argued, it is disingenuous

to assume that agents have some unlimited capacity for making binding commitments; commitment mechanisms, we claimed, must be built into the structure of the game. Nevertheless, if the state of nature *were* a Stag Hunt, or its equivalent, then contractarianism would succeed if it could be shown that such opportunities for commitment were or are available to agents and that a significant number would take advantage of those opportunities. Essentially, this is the argument of the current treatise.

First step: Is the state of nature a Stag Hunt? How do we answer that question? We might want to say that no individual model is an adequate representation of the state of nature, because the problems of the state of nature are variable. Sometimes agents will find themselves bargaining over resources that are both easily divisible and harvestable by one person, as in the Prisoner's Dilemma. Other times, the resources will spoil or completely go to waste if mutual fighting takes place, making both agents eager (too eager?) to strike a bargain. Finally, agents will encounter situations where resources can only be fully taken advantage of as a result of joint effort, and both will commit in a way that makes bargaining maximally advantageous to the other player.

The above discussion is parallel to Gauthier's [1986, p. 84] account of the perfectly competitive market as a "morally free zone." For Gauthier, "Morality arises from market failure." The state of nature is only a problem for rational agents because the perfectly competitive market cannot be realized. The real world difficulties of uncertainty, externalities, and transaction costs are what move the agents away from the optimal bargaining point that would be realized as a result of market activity and towards games like the Prisoner's Dilemma, Chicken, and the Stag Hunt. For example, if agents could be certain of the commitment of others, as in a perfectly competitive market, they

would have no problem finding the optimal bargaining point in the Stag Hunt. If externalities (fighting) did not consume resources, then agents could threaten and negotiate until the optimal bargaining point was reached in the Prisoner's Dilemma and Chicken games.

Since the world diverges from a perfectly competitive market in many ways, some of which we specified, it is clear that no game theoretic model will suffice on its own to capture these many differences. At best, perhaps it can be shown that processes running concurrently with the market – evolutionary dynamics – will tend to adjust the seriousness of fighting in the state of nature so that the situation will tend towards one of the models above the others. This is how we will argue for the Stag Hunt as the model of the state of nature instead of the other games. Evolution will restrict the bargaining space so that, at its edges, it points to the optimal bargaining solution. When this occurs, as in our third diagram, the state of nature will resemble a Stag Hunt, and the underlying dynamics will move agents to the cooperative and mutually beneficial equilibrium.

The state of nature, many of us have agreed, is more of a thought experiment, an account of what *could* have happened. For Gauthier [1986, p. 9], "Moral principles are introduced as the objects of fully voluntary *ex ante* agreement among rational persons. Such agreement is hypothetical, in supposing a pre-moral context for the adoption of moral rules and practices." Even though few of us have ever *inhabited* that pre-moral context, the argument is supposed to give us reason even now to abide by the constraints that we *would* commit ourselves to in the state of nature.

Gauthier admits that the difficult step in the argument is not in showing that agents would agree to certain constraints in a suitably terrible state of nature, but that they

would abide by those constraints when not in that situation. "Is it rational to internalize moral principles in one's choices," he asks [1986, p. 15], "or only to acquiesce in them in so far as one's interests are held in check by external, coercive constraints? The weakness of traditional contractarian theory has been in its inability to show the rationality of compliance." Thus, the problem lies in connecting the agreement that would be reached in a hypothetical state of nature with our real world conduct. This is the primary problem that, as a contractarian, Gauthier sets out to solve. By characterizing the state of nature in terms of real world evolutionary dynamics, it is not the primary problem for us.

Or, to put it another way, by allowing empirically testable, realistic evolutionary processes to sculpt the state of nature, our theory begins to move away from hypothetical agreement. The more we explain the state of nature in terms of our world, the more it becomes like our world. It does not have to be *exactly* like our world, but only similar to it in the aspects that primarily impinge on cooperative activity in some way. At some point, we will see ourselves reflected in the state of nature, the constraints agents would openly commit to hypothetically now merely implied. Since the state of nature is so similar to the real world, if we would commit to those constraints there (ex ante) then we probably would commit to them here (ex post) – or, at least, *someone* would make the commitment for us. This might sound unfair, but, as the next section will demonstrate, a contractarian argument does not have to restrict itself with considerations of fairness or impartiality in order to do the job it is supposed to do.

4. Impartiality & the Lockean Proviso

The title of Gauthier's book is Morals By Agreement, but this should not be taken to mean that he believes that anything agreed to is necessarily moral. To qualify as a moral principle, an agreed upon constraint must also be impartial, which really means it must be agreed to from an appropriately impartial situation. You may accept all kinds of rules under threat of torture, but for Gauthier, these would probably not qualify as moral rules, no matter how benevolent the result of the agreement.

Why the need for an impartial bargaining position? It is tempting to reply that impartiality is simply what morality demands, but moral contractarianism requires us to find non-moral reasons for restricting the initial bargaining position in this way. The state of nature must be partially civilized for rational agreement to take place. For Gauthier, gunfire and the threat of gunfire must be banished from the state of nature precisely because such coercive methods prevent rational agents from making commitments. The Lockean Proviso, then, is an impartial "precondition" for agreement.

As Gauthier argues [1986, p. 191-192], "Rational procedures yield a rationally acceptable outcome only from a rationally acceptable initial position." The Proviso – which "prohibits bettering one's situation through interaction that worsens the situation of another" [p. 203] – is Gauthier's way of constraining the state of nature so that the bargains people make within it are rationally satisfactory. His claim is that rational agents who made an agreement under the conditions of the Proviso would follow through on the agreement – that the Lockean Proviso allows agents to commit themselves.

We have agreed that agents cannot make commitments "on the fly" – that opportunities for commitment must be built into the game, and hence into the state of

nature. Does using the Lockean Proviso to limit the state of nature provide a satisfactory "commitment enabler", in just this way? This section will prove that it does not, because the Proviso does away with real world factors that make commitment both necessary and possible: violence and power imbalance.[8] This will show that you don't need impartiality to get commitment, so moral contractarianism can do without the Proviso. Finally, we will address the claim of Jan Narveson that the proviso can be had by agreement in the state of nature, after which it would act as a constraint on future interaction.

First, it must be admitted that the state of nature, hypothetical or not, must meet some conditions for agreement to proceed, so in this Gauthier is on the right track. If people only sought to worsen others, as many as possible, as much as possible, there would be little for them to bargain about. Even if you happened to be a strange mutant with preferences for things other than sadism, you would still be unable to bargain with everyone else. The problem in this case would not be with the choice procedure (utility maximization), but with the preferences fed into it, in particular the absolute preferences in others for your own suffering.

Thus, the first limitation on the state of nature is that most of its inhabitants be free from a preference for violence as an end in itself. This condition is fairly weak and historically quite well supported; that societies exist at all is an indication that people prefer food, shelter, etc. to violence and will inflict harm (only?) as a means to their procurement. In addition, even on his most charitable reading, Gauthier's constrained maximizer gains his advantage only from the widespread absence of "kingmaker"

[8] "Power imbalance" here is used to a refer to unequal bargaining power, no matter how "legitimately" the inequality came about.

strategies that would seek solely to undermine him.[9] But there is a wide gap between eliminating those from the state of nature who prefer to worsen others as an end in itself, and those who either have or are threatening to worsen others as a means to achieve many different goals. Those in the former group may safely be eliminated because they inhibit agreement completely, but those in the latter – that is, those credibly threatening violence against others – may be the ones who make such commitment possible. If the argument in this section is remotely correct, then contractarians will require in their state of nature individuals with a willingness to inflict harm on others to gain their own ends. At the very least, it will demonstrate that agreement with such individuals is possible, and that therefore their elimination is not a precondition of commitment on a social contract.

What is at issue is essentially the rationality of threat compliance. Acceptance of the Lockean Proviso is a prerequisite of agreement because, according to Gauthier, rational individuals would otherwise have no reason to comply with the object of the agreement once the threat was removed. "They may coerce and be coerced," he explains [1986, p. 198], "but they do not confuse coercion with cooperation." To illustrate this claim, he tells the story of a society of slave-masters that does away with its coercive instruments on the slaves' agreement that they will serve willingly. Supposedly, the agreement will leave both sides better off: the masters will be saved the cost of enforcing their rule (and can devote a small part of the savings to the slaves) and the slaves will be saved the suffering of frequent beatings. But, as Gauthier is correct to point out, after the threat of punishment is removed, the slaves no longer have reason to comply with their

[9] For example, consider a strategy that cooperates with straightforward maximizers and unconditional cooperators but never with constrained maximizers. Such a strategy would increase the utility of the first two strategies, perhaps offsetting the disadvantage they possess relative to constrained maximization.

earlier agreement. In our parlance, the agreement was not the same as a commitment, at least not once the whip was put down.

For Gauthier, rationality demands that agents only commit themselves to agreements made in the absence of coercion – that is, commitments made under the auspices of the Lockean Proviso. We have already argued that commitment is largely situational – that believable commitment is only possibly *sometimes*. Is it the case that such commitment opportunities are circumscribed to situations that satisfy the Lockean Proviso? No, certainly not.

To see why, consider Schelling's kidnapper example [1960, p. 43.] A kidnapper holds a wealthy heiress hostage. His ransom demand, while substantial to him, is fairly negligible for the woman. Once the ransom has been delivered, the kidnapper faces a dilemma, because if he releases the woman she will be able to identify him to the police, leading to his capture for sure. On the other hand, the police hunt murderers with greater intensity than they hunt kidnappers, so to kill the woman would increase the likelihood of his arrest as well (though not to the point of certainty.) While the woman promises to tell no one if he releases her, he knows that she will have no incentive to keep that promise after she has left his custody, and every reason not to keep it. If he is caught, she will get her ransom money back as well as a considerable amount of psychic benefit from knowing her kidnapper has been brought to justice. Unable to gain a believable commitment from his victim, the kidnapper has no choice but to kill her.

Compared to the feasible alternatives, this is a Pareto *inferior* turn of events. Both the kidnapper and his victim would prefer it if an opportunity for commitment existed for the woman. In fact, Schelling has just such a solution, suggesting that, "If the victim has

committed an act whose disclosure could lead to blackmail, [she] may confess it; if not, [she] might commit one in the presence of [her] captor, to create a bond that will ensure [her] silence." The kidnapper must believe the secret she tells him is at least as valuable to her as the ransom money and any psychic benefit she would receive from his capture. Only when this condition is satisfied can the secret become valuable to him as well, as a means to assure the reliability of the commitment. Commitment made, the kidnapper gains the money and his freedom, and the woman gets to keep her life.

The whole point of this example is that it is rational for the woman to make just such a commitment to the kidnapper even if doing so leaves him much better off than it leaves her. His possession of the gun and his willingness to use it creates an asymmetry in their relationship, but it is this asymmetry that allows her to commit to him in the first place. Arguably, she only needs to make that commitment because he kidnapped her, yet it also needs to be made clear that commitment is almost always a response to *some* other person's violation of the Lockean Proviso, or at least his threat to violate it. We will enter into "confederacy", Hobbes says, to decrease our own vulnerability. But other times, it will be in our interests to commit to those with whom we are especially vulnerable, but who may not be especially vulnerable to us. Gauthier would argue that it is irrational to make commitments of the second type, but there seems to be little difference between the two. In either case, the commitment is made only because a threat exists, and, again in either case, the terms of the agreement may not be "fair" to one side. For example, in an alliance between two innocent, but differentially vulnerable parties, the terms of the alliance might call for more sacrifice on the part of the weaker party

simply because he needs the alliance more.[10] If making an "unfair" deal can be rational when the threat comes from *without,* then surely it can be rational when the threat comes from *within.*

In his book, Gauthier distinguishes between broad and narrow compliance. A broadly compliant individual "is disposed to cooperate in ways that, followed by all, merely yield her some benefit in relation to universal non-cooperation." Narrowly compliant individuals are "disposed to co-operate in ways that, followed by all, yield nearly optimal and fair outcomes." [1986, p. 178] In the context of the above discussion, this means that narrowly compliant individuals differ from broadly compliant ones in that they will only comply with agreements made under the restrictions of the Lockean Proviso. Naturally, this excludes the agreement negotiated between the kidnapper and his victim.

Notice, first, that the result of narrow compliance in this case is Pareto *inferior* to the commitment reached by way of broad compliance. Gauthier would agree that sometimes narrowly compliant individuals must forgo certain opportunities to better themselves, so that broadly compliant individuals appear to have the upper hand in some situations. But as he would undoubtedly claim, a disposition to be broadly compliant is overall a bad thing, since it makes one a target for individuals like the kidnapper.[11] In

[10] Consider our bargaining example: the optimal bargaining point is only egalitarian, calling for equal sacrifice from both sides, because the parties are symmetrical in their endowments. Otherwise, it would favor the party with the least to lose if fighting occurred. This would seem to be the case even if a third party is the one threatening to fight. Surely Gauthier would not argue that it is irrational for the more vulnerable agent to forgo the bargain simply because it is not egalitarian! More importantly, since the optimal bargaining point corresponds with the mutually cooperative outcome, this can only mean that a "narrowly compliant" individual could not cooperate, leading to Pareto suboptimality. We will return to this point when we address Narveson's claims about the proviso.

[11] "For in so far as she is known to be broadly compliant, others will have every reason to maximize their utilities at her expense, by offering 'co-operation' on terms that offer her little more than she could expect from non-co-operation." [p. 179]

this, he may have a point. Having to murder the woman (even while getting the money) may be worse for him than not kidnapping her in the first place. In a society divided between broadly and narrowly compliant individuals, kidnappers will only have incentive to take hostages from the former part of the population. This means that, on the average, narrowly compliant individuals will be more successful than broadly compliant ones, at least in some respects. A population committed to narrow compliance could eliminate kidnappers completely; thus, narrow compliance is essentially one way of promoting and enforcing the Lockean Proviso. The problem is that this doesn't show that it is rational for any *individual* in the society to be narrowly compliant.

Narrowly compliant individuals reveal different preferences from broadly compliant individuals. The difference between the two dispositions, we must say, is in the way they evaluate situations, and not in the way those evaluations translate into action. If the woman in Schelling's example were broadly compliant, then, according to revealed preference theory, she would apparently prefer death to commitment to the kidnapper. This should be considered very strange. For example, suppose the kidnapper asks only for a dollar to spend before he dies of some incurable disease (thus minimizing the cost of making the commitment.) Would Gauthier still maintain that it is rational to die instead of giving him the money and making the commitment?

One might argue that the problem is that even if the kidnapper is only asking for a dollar this time, to give in would be to invite further kidnappings with escalating ransoms. However, the answer to this possibility is simply to give the kidnapper an opportunity to commit himself to silence (regarding the woman's revealed secret and broadly compliant disposition.) Perhaps having the woman perform some especially

heinous act with him, such that both would prefer the deed remain secret, could achieve this kind of double commitment. The point is that the attractiveness of narrow compliance is not universal, but circumstantial, and it would seem that the most rational agents (and indeed, most real people) would practice *selective* compliance, always considering the consequences of their behavior. If a mugger wants the change in your pocket, perhaps it is best for you to comply with his demand. It is not at all clear that this would increase the probability of your being mugged in the future.

Gauthier would claim that such compliance is not really *agreement* but only *acquiescence*. So it might be. But, as we have shown, in the state of nature, the bargaining zone might be very wide, leaving open many Pareto efficient opportunities, all of which are less "fair" than our proposed optimal bargaining solution. This means that acquiescing to an unfair deal can be a perfectly rational response if a player is able to make a credible commitment. And it is true that only the continued threat of violence guarantees compliance with such agreements – but so what? We are almost always motivated to keep our agreements because of what *would* happen if we did not.[12] This is why we continue to cooperate in the iterated Prisoner's Dilemma. This is why the slave-masters could have their agreement with the slaves, without fear of rebellion. All it would require is an understanding that the masters *would* pick up their whips again if any slave tried to rebel.

As one final gambit, Gauthier might claim that if the slaves were narrowly compliant they would prefer death to enslavement; if every last slave would rather die than serve, the argument would go, then the masters might as well release them, as killing them all would take a lot of energy and gain nothing. But consider the problem with this

[12] Binmore makes this same point.

kind of argument. For one, it should dispose the *masters* to a form of narrow compliance. The masters prefer the slaves to serve, but, if broadly compliant, prefer releasing them to having to kill them all for nothing. The slaves, if narrowly compliant, prefer their freedom to death, but death over enslavement. But the slaves only have reason to adopt this preference structure because they will gain their freedom from it. Narrowly compliant masters, recognizing this, would then have reason to alter their preferences so that they prefer to kill all the slaves instead of releasing them, thus taking away any reason the slaves might have to be narrowly compliant themselves.[13]

An objection that could be made is that once the slaves alter their preference structure to be narrowly compliant, they won't *want* to go back to being broadly compliant, even if it means their deaths. But the same could be said for the masters, who, having committed themselves to killing the slaves, would no longer want to back to preferring their release.[14] In any event, mutual narrow compliance seems to lead to more violence, not less, and certainly need not make the state of nature more civilized along the lines of the Lockean Proviso. Because the slaves are especially vulnerable, they, like the woman in Schelling's example, may need to commit themselves to a course of action that is much more beneficial for the other party, but still Pareto superior to narrow

[13] Is it accurate to describe the slave masters in this case as "narrowly compliant"? Gauthier builds fairness into his definition of narrow compliance, but perhaps his reasoning here is circular unless he really refers to a perception of fairness. If the rationality of narrow compliance gets you the Lockean Proviso (because rational individuals wouldn't make agreements formed any other way), and the proviso is what ensures bargains are fair, then isn't it just circular to define narrow compliance in terms of fairness? For better or for worse, this treatise is concerned more with the *perception* of fairness, which is not directly susceptible to an argument like this.

[14] A paradox lurks here: At T1, the slaves prefer service to death and the masters prefer freeing the slaves to killing them all. At T2, the slaves become narrowly compliant and prefer to die instead of serve. At T3, the masters have a choice to make. Right now, they prefer to release the slaves. But they know that *if* they become narrowly compliant, then they will prefer killing them. If they change their disposition, then they will end up killing them, which is something they don't prefer *now*, but will prefer at T4. So narrow compliance is irrational at T3 but will become rational at T4. Is this what people mean when they talking about rationalization after the fact?

compliance (that is, death.) Such commitments are admittedly unfair – and may, as Gauthier suggests, offer some parties terms that are little better than non-cooperation – but that just demonstrates that we need not obtain fairness to have agreement.

This is all to undermine the rationale for narrow compliance. Gauthier's claim seems to be that situations must satisfy the Lockean Proviso before a narrowly compliant individual would accept agreements made within them. However much this may be true, its relevance still rests on the rational superiority of narrow compliance. We have argued that narrow compliance is not rationally superior, and may in fact be self-referential and even incoherent (if he is narrowly compliant then I will be narrowly compliant but then he will be broadly compliant so I will be broadly compliant, etc.) But if rational agents can make commitments even in coercive situations, the sole non-moral justification for an impartially restricted state of nature vanishes. Some might argue that the social contract must be made under impartial circumstances (i.e. behind the veil of ignorance) to count as moral, but injecting a requirement of impartiality for moral reasons defeats the contractarian's goal of deriving moral principles from premises devoid of moral content.

Gauthier claims that the Lockean Proviso is a precondition for rational agreement. But what if the proviso itself was the outcome of agreement and then acted as a restriction on all future interaction? This is essentially Narveson's assertion. One of the conditions of the Hobbesian (moral) state of nature, he argues, is a broad equality of vulnerability. We can all be killed and have our lives worsened in a myriad of ways. No one wants to be worsened, but it only becomes rational for others to agree not to worsen us if we agree not to worsen them (Hobbes' second Law of Nature.) Thus, the proviso represents a sort of social minimum, in that we can expect pretty much everyone to

accept it, since those who do not will remain in a state of war against us (Hobbes' first Law of Nature.) Of course, an implication of the Lockean Proviso is that we keep our agreements, since without it an agreement to keep the proviso is not worth much in the first place (Hobbes' third Law of Nature.) After accepting this general social minimum, individuals can make other agreements with each other and proceed on that basis towards a more civilized society.

In the state of nature, we might be equally vulnerable and agree to something like the Lockean Proviso. We might even say, "I agree to be bound by the proviso now, even if I find myself far less vulnerable than some of you in the future (if, for example, I happen to be in a position to enslave you.)" This kind of agreement makes a lot of *sense*. Imagine if instead I declared, "I agree to be bound by the proviso – but only as long as I'm as vulnerable as the rest of you." Although we could all *say* that and perhaps even commit to it, this latter agreement would not get us very far from the state of nature. Even Hobbes' Foole will cooperate when he can see that the alternative is death; the challenge for him is sticking to the proviso when he can exploit people without fear of reprisal. To solution to the challenge must be to show that it is both feasible and rational to make a commitment of the first type as well. This commitment then would serve as the social minimum.

The idea of an agreed social minimum, a commitment to a basic constraint on behavior, is very attractive, and is the main reason we are dealing with Narveson's argument. This treatise takes issue with the equality of vulnerability assumption, arguing that people are not *essentially* equally vulnerable, at least not in the sense required. Certainly, some kind of equality of vulnerability exists, but it isn't enough from which to

deduce the Lockean Proviso as *the* social minimum. There are basically two ways of reading "equality of vulnerability" and both seem to be consistent with Hobbes' view of the matter. On one hand, equality of vulnerability can mean that everyone *can* be killed, worsened, etc, which I take as uncontroversial. On the other hand, equality of vulnerability can mean that everyone *can* kill, worsen, etc everyone else, and this *is* controversial, or at least should be. The first kind of equality does not get us to the proviso, and the second kind requires us to restrict the state of nature even more than Gauthier would desire if the Lockean Proviso were to be reached.

First: yes, everyone can be killed. We may even take this to mean that, in the state of nature, for any person, there is at least one other person who is in a position to kill him. But *I* may not be the one in that position, for various reasons. So a rationally acceptable social minimum may not prohibit worsening *all* people, but only those who stand a good chance of being in a position to worsen *me*. Obviously, such selective restraint is very different from the impartiality the Lockean Proviso requires. It will be argued that, while others may not be in a position to worsen me right at the moment, they will be in that position sometime in the future. Still, it seems unlikely that all others have an *equal* chance of occupying that state, so perhaps the amount and direction of my restraint should be directly equivalent to the likelihood that others will be in a position to harm me and to the harm they are capable of inflicting upon me. Again, this is selective and very much different from the universal, unequivocal constraint the Lockean Proviso demands.[15]

[15] It also pushes us from the libertarianism that Narveson prefers to a strange sort of utilitarianism: the average person will be, on the average, in a position to harm other average people most of the time. Statistically, then, most constraint will be delegated in a way to minimize harm to the average person (definitely another paper topic here.)

At the same time, it is not clear that the ability to harm someone else is a reflective characteristic (that is, it is not always true to say that if I am in a position to harm you then you are in a position to harm me.) Suppose, for instance, that I am part of a mob descending on your house. As part of the mob, I am in a position to hurt you, but since I am one of many people, you may not be in a position where you will be able to hurt me. Even if you lash out randomly at the mob, if the group is large enough there may be little chance that I will be the one to suffer. Again, my restraint should be proportional to how dangerous you are to me. On the other hand, if I enforced the proviso *against* the mob (by, for example, refusing to participate or even helping you), then I could become a target myself. And the mob is in a much better position to damage me than you are.[16] Within the mob, a constraint like the Lockean Proviso probably *is* rational, just as it is rational not to worsen the people who live next door and have a fairly large assortment of automatic weapons in their basement. But this is no different than the Foole's selective restraint we examined earlier.

Thus, it is unclear how we can get from the first type of equality of vulnerability, which is empirically supported, to the Lockean Proviso, or perhaps to any constraint at all. It is also unclear how we can get from the first type to the second type, which, in any event, seems empirically untenable. Still, the state of nature is hypothetical to begin with, or so it is according to this argument, and a lack of evidence for the second kind of equality is not necessarily a reason to remove it from consideration entirely. But if it can be shown that, even when equality of vulnerability obtains, it remains irrational for some

[16] This raises an obvious question: what's to prevent any of *us* from becoming the victim of a mob? Or of one mob becoming the victim of another mob? However, it's not clear if the rational response to this possibility is to accept a prohibition against mob violence, or simply to cling to the biggest, most well established mob – perhaps the mob other people *expect* you to cling to?

individuals to commit to the proviso, then some other condition must be required to make the Proviso rational. That condition, we will suggest, is much like a veil of ignorance, taking us far from the Hobbesian state of nature.

Paul Viminitz [2000] suggests an inventive scenario in which a plague threatens the entire population and can only be cured with a drop of a certain person's blood. This person, Jones, is unwilling to part with the drop for any price.

> The libertarian cannot discount the possibility of the scenario with the assurance that we will cross that unlikely bridge when we come to it. For the question before us is not what we might do then, but whether we should put in place now institutions that will ensure that Jones doing as she pleases with her own blood cannot be interfered with.

For the agents in the state of nature to commit to the Lockean Proviso is to create just the sort of (moral) institution that will serve as a fundamental constraint on all interaction in the future. If the commitment to the proviso were *genuine*, then the agents would not be able to interfere with Jones and, presumably, the world would perish. If the agents in the state of nature knew that a plague like this one was on the horizon, would they commit to the Lockean Proviso, even if they were equally vulnerable to *each other*? If not, then equality of vulnerability, even of the second type, isn't enough to get to the Lockean Proviso.

There are at least two ways to deal with Viminitz' argument. On one hand, we can argue that as events like killer plagues are extremely rare, is it really much of a failing if our principle can't handle them adequately? Isn't it enough to say that commitment to the Lockean Proviso is rational almost all the time – certainly more often

than any other general principle could claim to be? Our second response to Viminitz may be to deny the existence of people like Jones. Certainly, Jones' existence is *logically* possible, but how likely is it that someone would be unwilling to lose a drop of blood for *any* price? Hence, Viminitz' argument against libertarian contractarianism requires the conjunction of two extremely unlikely events. Our commitment to the Lockean Proviso could be selective, but not *very* selective; we would only be a little bit like the Foole.

Thus, the other condition underlying the rational adoption of the Lockean Proviso – or one of the other conditions – is that the universe be fairly free from events like the plague and also without very many agents like Jones. But what *is* an event like "the" plague? What is an agent "like" Jones? It seems we are again close to the Foole's selective restraint, where we commit to the Lockean Proviso only as long as we cannot gain more by ignoring it.

Viminitz asks, "If these hard-wrought protections can be justifiably dismantled in the face of a threat to subsistence, why not do so in the face of any circumstance in which someone expects to gain more from dismantling them than from maintaining them?" This makes it sound as if our selectivity in the application of the Lockean Proviso is *arbitrary*. But there may be very good reason to believe that individuals like Jones, who would be unwilling to part with a drop of blood for any price are very unlikely to exist, and that events like the killer plague are even less likely. The question here would be empirical, or real-world, and the rationality of the proviso as the social minimum – the rationality of *any* social minimum – depends on the answer. Who we are, where we live, and where we've been, makes all the difference.

To illustrate this point, consider what would probably happen if our scenario were to really take place in an ostensibly libertarian world. Millions of people are at risk and Jones' blood holds the cure. By our intuitive, non-rational, and specious interpersonal comparisons of utility, we know that the harm inflicted on Jones if we take his blood is very minimal, no matter how he protests. Strapping him to a table, we stick a needle in his arm, and violate the Lockean Proviso. Afterwards, utilizing our inadequate interpersonal calculations, we might try to find a way to repay Jones for his "sacrifice" – but I am not altogether sure the scenario would have played out differently had the cure for the plague involved killing Jones instead of just taking a bit of his blood.

As a general social norm, the Lockean Proviso lacks stability to the extent that events like plagues, meteors, and people like Jones exist in the world. If some lucky agents *did* escape Viminitz' plague, how soon would they be to agree to the Lockean Proviso as the new basic rule of their potential civilization? Or would their past experience lead them to be more selective in the terms of the social contract? "No worsening others to better yourself – except in case of plague." What we intend to show is that such selectivity can be generalized; rational agents, faced with a state of nature where mistakes and ignorance and plagues are permitted, will realize that their best shot for commitment lies in the application of a fairly *non-rational* strategy. They will realize that, to a great extent, they ought to commit themselves to the imitation of others, especially those of previous generations. That is, they will realize that they ought to be conservatives.

5. Where We've Been

"Given certain motivations, and placed in an initially inefficient situation, subject to certain limitations to ensure the impartiality of the contract, agents would be rational to agree to certain constraints on their behavior." Such was our general characterization of the contractarian enterprise. In the previous sections, we used Gauthier's Morals By Agreement as a specific example of this contractarian model and criticized it in several ways. In preparation for the constructive argument of this treatise, the list below summarizes those criticisms.

First, contractarianism requires not just agreement, but commitment: As we use the term, a commitment is a binding promise to employ some strategy rather than another. When an agent so binds himself *ex ante* (in what we might call the commitment stage of interaction), he will *certainly* carry through on his promise in the *ex post* situation (the implementation stage.)

Second, rational agents cannot make commitments on the fly: Gauthier's notion of a disposition requires his agents to be able to infallibly commit themselves to one course of action of another. This requirement conflicts with the tenets of rational choice. Instead, we argued that commitment mechanisms, or *enablers* must somehow be built into the game being played.

Third, the state of nature is inefficient relative to some "optimal bargaining point": However, unlike Gauthier, we do not argue that any game theoretic model can capture the inefficiencies of the state of nature because they come in several different forms. Instead, by connecting the models to an underlying bargaining game, we

established a more "dynamic" concept of the state of nature. The appropriate model depends on the shape, size, and scope of the bargaining zone. This allows us to think of a state of nature that can *evolve smoothly* as these factors change.

Fourth, commitment does not require impartial constraint: there is no rationale for limiting the state of nature according to the Lockean Proviso. As Narveson argues, agents can commit to a social minimum from a completely "uncivilized" state of nature. However, we proposed that stable commitment to the Proviso requires that the state of nature be limited in other ways. Our final claim was that in a state of nature subject to mistakes, random shocks, and other "stochastic phenomena", agents would generally commit themselves to the imitation of the behavior of others. In the next section, we will prove this claim by using a computer simulation that hopefully captures a few of the relevant aspects of social interaction. Below we list some of what the program is intended to demonstrate.

First, that cooperation can be a fairly successful strategy in a non-stochastic state of nature, where all agents apply an identical, fully informed learning procedure: using the underlying bargaining game as our model, we will show how fully cooperative agents succeed over agents who are only partially cooperative (players who cooperate only some of the time) and players who are never cooperative.

Second, that when agents are permitted to make mistakes and behave in an ignorant manner, cooperation becomes less stable: The rational learning procedure agents use to decide who to interact with is insufficient to counter error and ignorance and keep the population fully cooperative. Some other dynamic is required to overcome to problems of ignorance and error.

Third, that introducing an "imitative" dynamic to the model solves the problem: agents react to nearby strategies and adopt them as their own. It will be shown that the *range* in which agents are allowed to perceive the behavior of others makes a difference. Having more complete information does not necessarily facilitate coordination on the optimally cooperative equilibrium. Instead, a certain amount of narrow-sightedness and conformity is actually better for cooperative agents.

6. The Simulation

"Ten strangers find themselves in a new location. Each morning, everyone wakes up and chooses someone to visit. Initially the process is random, but the visits result in interactions, and a pleasant interaction with someone makes it more likely the visitor will visit that person next time." It is through a process as simple as this one that Skyrms [2004] argues *some* kind of cooperative activity can come about. Describing the application of this procedure to the Stag Hunt game, he argues, "Stag hunters quickly learn to visit only other stag hunters. This is not surprising. Visits to hare hunters are not reinforced, while their visits to stag hunters are." This eventually leads to a situation where stag hunters visit only other stag hunters, while hare hunters are consigned to a "second-rate existence", playing only amongst themselves.

This is an adequate outcome for the contractarian, to be sure, and a way of "civilizing" the state of nature so that it looks very much as if it were under the auspices of the Lockean Proviso. In this world, certainly, the agents who settle for the optimal bargaining position do very well indeed. What we intend to show is that, placed in a world where mistakes can be made, such an open-minded form of associative learning succumbs to strategies that initially appear to be very irrational. When agents imitate these strategies with sufficient frequency, it becomes advantageous for other agents to do so as well.

In the simulation designed for this treatise, twenty-four agents are situated in a circle, initially interacting with each other at random. Rather than having them play Prisoner's Dilemma, Chicken, or Stag Hunt specifically, the agents will play the

underlying bargaining game, known as the *Nash demand game* [Axtell et al. 2001, p. 194.] The basic form of this game has three strategies, shown in Figure 6.

	High	Medium	Low
High	0,0	0,0	70,30
Medium	0,0	50,50	50,50
Low	30,70	30,50	30,30

Figure 6: Nash Demand Game

The reason the Nash demand game has been selected instead of some other game goes back to our more holistic account of the state of nature. First, unlike the Prisoner's Dilemma, it is difficult to describe any of the strategies in the Nash demand game as *essentially* moral or immoral. Arguably, it is only what others expect you to do in the game that makes one strategy stand out more than the others, as we shall see. Second, the Nash demand game can encompass any of the other games, depending on the *expectations* of the agents. For example, suppose I expect my opponent to bid high. My only rational alternative in that case is to bid low, leading to payoffs of (30,70.) Bidding anything else would have led to nothing for the both of us (or so I would believe.) My expectations led me into a situation that had precisely the same structure as a Chicken game. In our vernacular, expectations (shared or otherwise) *restrict* the bargaining space.

Thus, the important issue is not so much in defining the rational thing to do in any *particular* game, but in describing the process by which the expectations that create the game are formed. As in our previous example, what goes on to create the game may in many cases also commit one or both players to a course of action. We shall therefore call the underlying bargaining game that potentially leads to such commitments the *commitment* stage of the formation of a social contract. We shall call the strategic situation that develops once those commitments are in place the *implementation* stage,

since that is when agents act on the commitments they may have made in the past. Our simulation is a model of the commitment stage; we are interested in examining the formation of shared expectations – what Robert Sugden [1986] would call conventions. These expectations can drastically influence the terms of the ensuing social contract, erecting a social minimum that will feasibly be very different from the Lockean Proviso.

Skyrms' associative learning procedure is a basic way of understanding the formation of expectations. Each agent keeps a record of his past interactions with other agents. This memory allows an agent to associate other agents with one of the three strategies. Essentially, the more often a particular strategy is used against an agent, the more the agent will expect that strategy to be used again. Thus, if paired with an agent who has bid high ten out of the last twelve interactions, the player will be very likely to bid low. This sort of learning procedure seems as highly rational as any that could be employed in a computer program.

Aside from the strategies derived from memory, agents also differ in color and in "status." As color is randomly distributed among the agents, it cannot have any meaning, at least initially. However, the "status" of an agent is either liberal or bigot. This determines how the learning procedure is applied. When a liberal agent is matched with another player, he uses his memory of his interactions with *all* players to determine his choice of strategy. In contrast, bigots only consider their interactions with agents of the same color as the one they have been matched with when forming an expectation.[17] Since color is randomly distributed, one might expect bigotry to be an irrational strategy,

[17] The notion of distinguishing agents according to an arbitrary characteristic was initially inspired by Axtell, Epstein, and Young's paper [2001.]

like an agent deliberately ignoring at least some of the information available. This is an assumption our simulation will eventually test.[18]

Our initial simulation will be without both mistakes and bigots; it is designed simply to show what kind of expectations (and hence what social contract) we might expect to arise from "perfect interaction." Even here, however, observations do not reveal an unequivocal triumph for cooperation on the optimal bargaining position. After an initial shakedown where strategies fluctuate wildly, an equilibrium is reached. We can observe the equilibrium because, at that point, expectations have become so settled that very few shifts in strategy occur. Basically, everyone knows what to expect from everyone else. Figure 7 depicts how the strategy distribution typically looks after one thousand iterations. There is always a good mixture of high bidders and low bidders; the medium bidders, while outnumbered, tend to score above the average, and usually above most of the high bidders.

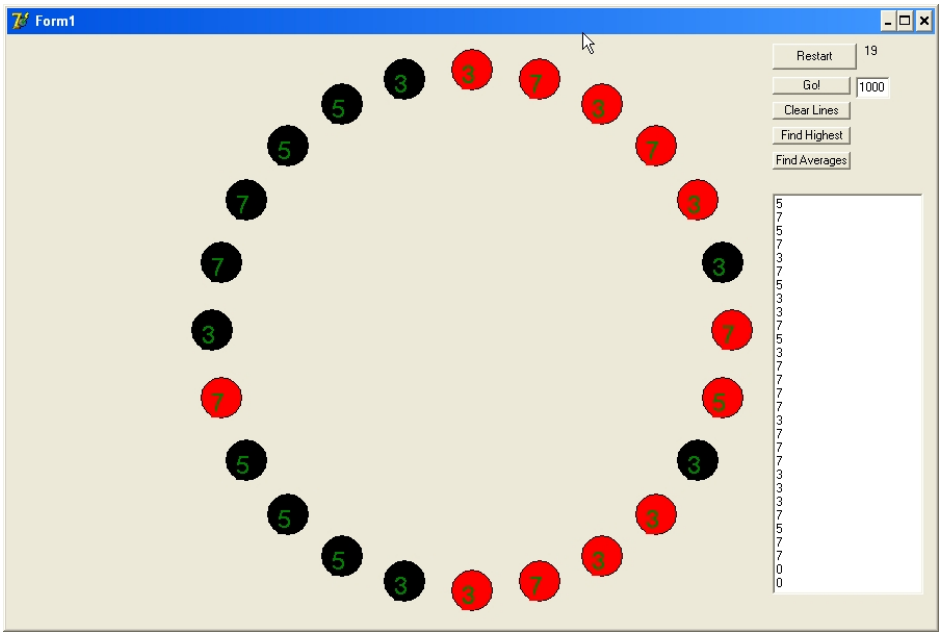

Figure 7: No Mistakes, No Imitation, and No Bigots

[18] I hereby designate this method of testing our assumptions about cooperative behavior and evolutionary dynamics through computer simulation, "cybernetic sociobiology."

All of this is only marginally interesting. What *is* interesting is what happens when we randomly distribute some bigots into the population. As expected, the bigots do poorly, *but so does everyone else*. This seems to be because liberals form inaccurate expectations about the strategies of bigots. When a liberal meets a bigot of the opposite color, the liberal uses his entire memory as input for the learning procedure, but the bigot only uses part of his. Because of this, liberals tend to have similar memories, and so tend to be right in their expectations about other liberals, but tend to be wrong when they interact with bigots.

While bigots force absolute payoffs to be lower, they also allow medium bidders to be dominant in the population. A typical run of the simulation is shown below.

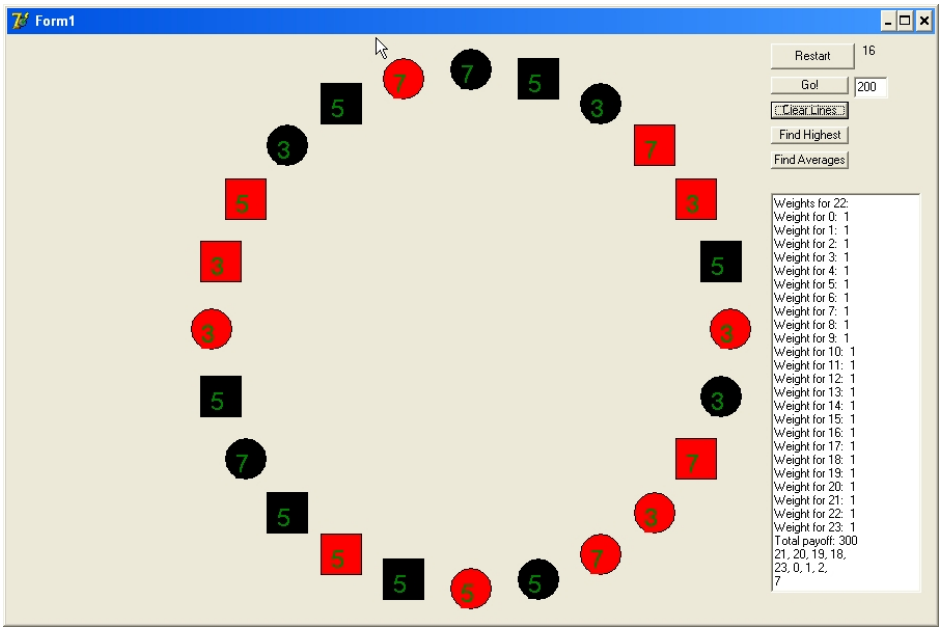

Figure 8: Bigots, but No Mistakes or Imitation

The bigots are the square shaped agents. As you can see in Figure 8, bigoted medium bidders have grabbed a significant portion of the population, especially the black variety. It seems that bigots create a sort of self-fulfilling prophecy: by acting as if color was an important predictor of behavior, they turned it into just such a predictor. These shared

expectations behave as coordinating conventions; they allow the medium bidders to survive when otherwise less scrupulous (but also less ignorant) agents would have wiped them out.

Next, we will allow the agents to make mistakes. Or, less pejoratively, we will allow them to *experiment randomly* sometimes, instead of slavishly adhering to their learning algorithm. About ten percent of the time, an agent will ignore his memory and make a bid at random. What's surprising when we do this is that low bidding agents begin to look the most successful! Agents who often bid low make small amounts consistently, no matter what associations they form. They only lose out in the long run because high bidding agents form the right associations and begin to exploit them. But mistakes hamper the ability of high bidders to form accurate memories. Thus, high bidders tend to lose out to the more conservative low bidders.[19]

However, allowing mistakes does not get us anything like an optimally efficient social contract. Low bidding agents do *relatively* well, but it is important to remember that two low bidders let forty percent of the "cooperative surplus" go to waste. The associative learning procedure, while rational and effective in some situations, is not enough to coordinate individuals on a social contract that is both fairly cooperative and efficient. What else do we need? The missing piece to this puzzle comes from Skyrms' work. Skyrms shows that cooperation can be achieved even in a repeated single shot Prisoner's Dilemma with the application of an "imitate the best" dynamic (henceforth known as the imitation dynamic.) Essentially, after each round, agents look to their "neighborhood" and adopt the characteristics (the strategy, status, etc) of the most

[19] Or, to put it another way, in a very chaotic world, it is better to play a simple coordination game (dual low bidders), then a game like Chicken (medium bidder and high bidder.)

successful player in that group. The imitation dynamic can lead to cooperation in many other games, so why not this one?

Every agent in our circle has at least two neighbors, to his left and right. Initially, the neighborhood will be restricted to only those neighbors, but in principle it could be extended to take into account the strategies of the four, six, eight, etc surrounding players. However, as we will show, a large neighborhood may actually *inhibit* cooperation; again, a certain amount of ignorance may be necessary to achieve the most attractive outcomes. After each round, an agent looks his neighborhood and, if any player in the group is more successful (in terms of total payoff) than he is himself, the agent will adopt both his strategy and his status. Potentially, then, liberals can become bigots and bigots can become liberals.

Watching the simulation for the first thousand iterations, it is impossible to miss how much more the population moves with the addition of the imitative dynamic. Sometimes, strategies that are doing poorly will begin to gain favor with the rest of the population, like the few low bidders in a society full of high bidders. Of course, as the number of low bidders increase, both high and medium bidders begin to do better. In short, imitation seems to severely restrict the stability of certain equilibriums. Except, perhaps, for the one shown below. Time and time again, this is the equilibrium the population settles on, even if it takes more than a thousand iterations.

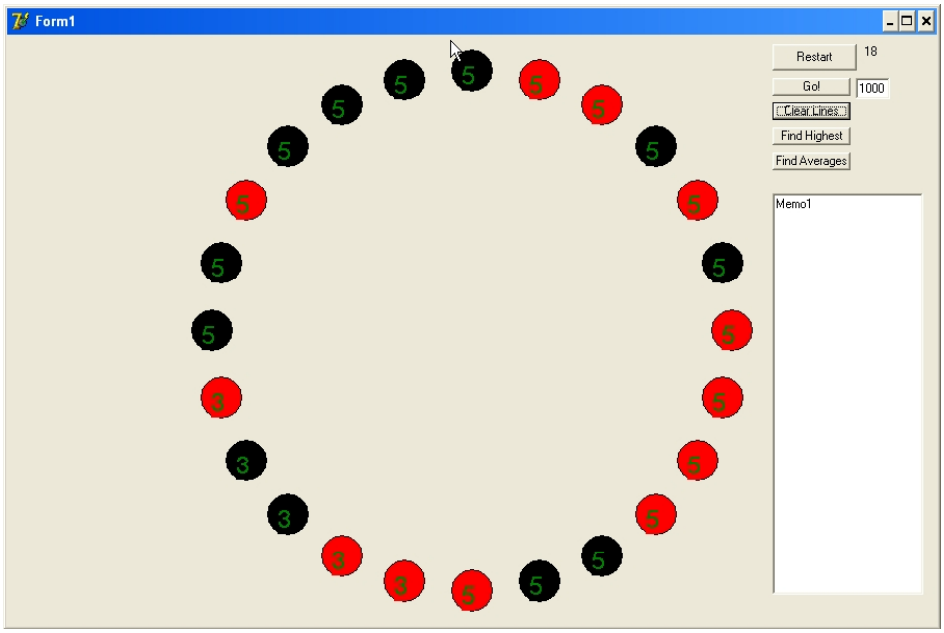

Figure 9: Introducing Imitation With Small Neighborhoods

In Figure 9, we have not only the triumph of the medium bidders, but also the triumph of liberalism. As the high bidders have been completely eliminated, the low bidders can subsist in the population, although probably if the simulation were run for longer they too would assimilate into the optimally cooperative, optimally efficient culture.

This is what happens when the imitation neighborhood is restricted to just the two agents immediately to the right and left of every player. What happens when that neighborhood is extended? Sometimes, the outcome is the same as above: liberal medium bidders take over almost the entire population. But this occurs less frequently than it does when the neighborhood is fairly small. For example, Figure 10 shows what often happens when the neighborhood is extended to six agents (three on either side.)

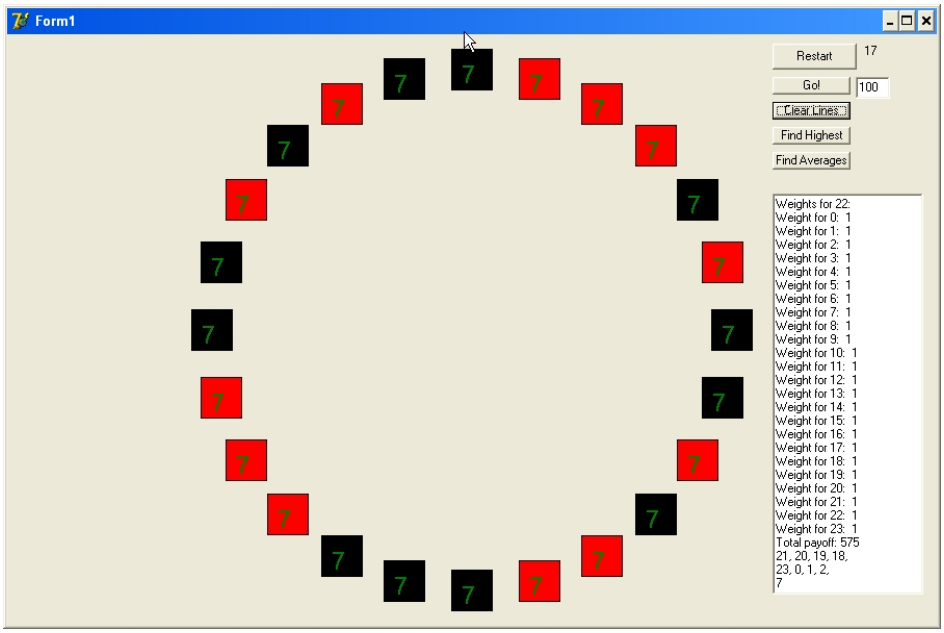

Figure 10**:** Imitation With Extended Neighborhoods

This would seem to occur when the imitation neighborhood is quite extensive and large swaths of agents simultaneously decide to imitate a few scattered high bidders (who are doing well, being in the minority.) Suddenly, the entire population can shift into high bidding mode, which is a fairly bad thing. While on any given round the associative learning procedure will direct agents to bid low, the imitation dynamic will quickly bring them back to playing high again. This is a situation one might almost want to call *extinction* for the population. This should remind us that imitation is still a fairly non-rational strategy to adopt, even if it leads to cooperative outcomes.

7. Summary of Findings

In slightly more diminished neighborhoods, medium bidding usually wins out, but liberalism often does not. There seems to be an optimal level of ignorance for agents to maintain. Shrinking the imitation neighborhood is one way of preventing players from having awareness of everything that is going on in the world, while bigotry is another. Finally, the memory size of the agents also plays a role in the equilibrium that develops. When memory is short, so the agents only remember their last ten encounters or so, medium bidders can still take over the population, but only if they are also bigots. If memories are extremely short, then even after several thousand iterations no equilibrium may be found. The population seems to shift cyclically. Very long memories eliminate not only bigots, but also medium bidders.[20]

While these observations are interesting, they are difficult to connect in any systematic way. The expectations that most facilitate cooperation seem to form in at least a partial vacuum of ignorance. A stable, efficient social contract will work within that vacuum, relying on expectations that have been formed through experience to coordinate individuals effectively. A rational agent coming into a society has no (rational) choice but to abide by those expectations. As in our simulation, if most of the red players are bidding low, and you are a red player, it then becomes in your interests to bid low as well; it becomes, for you, a *commitment* to bid that way. Collective commitments

[20] Here is one possible explanation: If the strategies of other agents are fairly unpredictable, high bidding is the riskiest strategy, as it restricts positive interactions only to those that occur with low bidders. However, these risks can be eliminated if agents are allowed to keep long records of their past interactions. In that case, chance fluctuations in strategic behavior can begin to normalize out, revealing if there is a reliable predisposition for a proportion of agents to bid low. And with reliable information, liberals have the advantage over bigots.

represent norms of behavior for that population, locking agents into Prisoner's Dilemmas, or Chicken games, or Stag Hunts.

If the simulation demonstrates anything, it is that even when starting from a relatively equal position, where agents are potentially vulnerable to all other agents, where everyone potentially has access to the same information, the most optimally cooperative norms are not guaranteed to develop. Instead, mistakes, bigotry, and localized interaction can lead to norms that, for example, give seventy percent of the advantages of the contract to some agents simply on the basis of arbitrary characteristics (i.e. color.) In the implementation stage of the social contract, these widespread expectations are what *determine* the rational course of action.[21] Thus, it would seem to be inaccurate to describe the Lockean Proviso as the *only* norm that agents would commit themselves to, even starting from a position that was equal in all the relevant ways.

[21] An example: suppose that, in the commitment stage, expectations have been set up for red players to bid low and for black players to bid high. As already argued, when a black player is matched with a red player under these circumstances, the game they end up playing is Chicken. The same expectations that *create* the Chicken game also determine what it is rational for each player to do in the game: the red player should cooperate and the black player should defect.

8. Conclusion

Bruce Ackerman [1980, p. 337] makes a general critique of contractarianism. He argues that contractarians of all types:

> …Want to convince us to approach the problem of justice as if we were (1) some hypothetical person with a particular set of preferences confronting (2) some hypothetical situation that forces us to choose among a number of options open to us. Once he has decided upon a proper characterization of (1) the ideal chooser and (2) the proper choice set, the contractarian's argument is straightforward: given (1) and (2), he wishes to demonstrate that choosers will reject certain policies within the choice set in favor of other policies open to them. It is these policies that will be inscribed in the social compact that is to structure subsequent social interaction between the parties.

Notice that, aside from omitting the commitment requirement, Ackerman's presentation is close to our own. The problem with contractarianism so stated, he claims, is that by carefully choosing his hypothetical (1) and (2), the theorist can ensure that his hypothetical agent will agree to anything he desires. This argument is not so new: we saw it before when we addressed Viminitz' claims against libertarianism. Part of Viminitz's case is that the contractarian can only get agreement on libertarian principles if he assumes that his agents value liberty above everything else. But to impute the selection of agent and initial position with certain characteristics solely to derive a particular set of moral principles seems question begging.

Both Ackerman and Viminitz (whether they know it or not) are looking for a non-*moral*, but also a non-*arbitrary* way of selecting circumstances and agents, such that

those agents under *those* circumstances would commit themselves to certain *moral* principles. It's the apparent capriciousness of the hypothetical contractarian's choice of agent and state of nature that makes people like Ackerman suspicious, as well it should. Here we will show that the choice need not be arbitrary, if indeed we have a real choice in the first place.

We have essentially argued that the expectations of agents, past and present, shape their cooperative opportunities. Social reality is not *just* a Prisoner's Dilemma or Stag Hunt; the expectations inherited from tradition can lead agents of a certain type to play a simple coordination game with each other and a Chicken game with everyone else. Moreover, as there is no non-moral reason to exclude these possibilities from consideration, the theorist ought to incorporate them into his model of the state of nature. We claimed the best way to do this is to understand the state of nature as a bargaining game in which the bargaining zone can be restricted according to the beliefs – well informed or otherwise – of the agents. Indeed, we showed that a certain amount of ignorance and (apparently) mindless imitation could stabilize a population around optimally cooperative norms.

Thus, our response to Ackerman is that, by and large, the traditions we have *actually* inherited function to form coordinating expectations in the social game we really are playing. In this way, our traditions are not arbitrary and actually make up an essential parts of the decision making procedure any rational agent would employ to solve a strategic situation, like a Stag Hunt. In reply, Ackerman might claim that we are not being impartial or neutral by privileging tradition over other values like equality. But inherited expectations might very well imply a more equal distribution of resources; we

know this because, as Binmore delights in pointing out, many less technologically advanced cultures are at least semi-egalitarian. What we inherit from our ancestors is not *necessarily* capitalism.[22]

And if we strip our rational agents of the cultural hardware they would *normally* use to solve strategic situations, what then? How *do* you decide whether to cooperate or not in a situation like the Stag Hunt? The only honest answer is that you look at what other people are doing and what they expect from you. There is no way to rationally solve the Stag Hunt or Chicken *except* by reference to the norms that are *actually* in operation. Thus, in at least some cases, tradition is not arbitrary but *essential* for rational agents, and as such we have good reason to use it as a guide when we select the characteristics of our agents and our initial bargaining position. Our computer models simply described *how* something as ambiguous as tradition could serve as a coordination device under most circumstances. Using tradition for this function worked best when agents at least partially bypassed their rational decision-making procedures and simply imitated those who had shown themselves to be the most successful in the past.

This conclusion is more radical than it may first appear. For if our model of the state of nature really does represent prototypical (if primitive) human interaction – and we have tried to make it fit at least in the barest detail – then the instinct to imitate must have developed *in conjunction* with our capacity for cooperative behavior. This instinct, it must be speculated, lies underneath rationality, informing our utility functions in the

[22] However, there might be good reasons to think that capitalism is more *likely* to develop than egalitarianism as a society progresses. Why not a libertarian theory of history?

same way the expectations formed in the bargaining game helped determine the rational course of action in the implementation stage of the social contract.[23]

Thus, we are, in some sense, already deeply engaged in the business of imitating each other; a contractarian theory that took as its starting point agents with a predisposition to imitate *in addition to or instead of* a capacity for rationality would have an admirable chance of coming up with moral rules that real people would commit themselves to. This, then, is essentially our theory: that moral principles survive mainly through the imitation of others, that we expect others to imitate[24], and that in cases where this expectation is not widespread, the rules no longer bind. In that case, we default back to the "state of nature", the bargaining game – in which, as will be recalled, no strategy can be described as inherently moral or immoral – until such time as new expectations arise, forming the basis for new commitments and hence a new social contract.

[23] More research is needed here, especially experimental studies. E.O. Wilson's Genes, Mind, and Culture provides a basic framework in which to understand how an imitative disposition could develop through natural selection. Consider, for example, the idea that Wilson's "epigenetic rules" could act as constraints on utility functions.

[24] Not that we *know* we expect others to imitate.

BIBLIOGRAPHY

Ackerman, Bruce. Social Justice In The Liberal State, New Haven: Yale University Press, 1980.

Axtell, Robert L. and Joshua M. Epstein and H. Peyton Young. "The Emergence of Classes in a Multi-Agent Bargaining Model" in Social Dynamics. Washington, D.C.: Brookings Institute, 2001.

Binmore, Ken. Playing Fair, Cambridge: MIT Press, 1994.

Boucher, David and Paul Kelly. "The Social Contract And Its Critics: An Overview" in Boucher, David and Paul Kelly, eds., The Social Contract from Hobbes to Rawls, New York: Routledge, 1994.

Gauthier, David. Morals by Agreement, Oxford: Clarendon Press, 1986.

Gauthier, David. "Morality, Rational Choice, and Semantic Representation: A Reply to My Critics." *Social Philosophy Policy* 5:2 (1988).

Lumsden, Charles and E.O. Wilson. Genes, Mind and Culture: The Co-Evolutionary Process, Cambridge, Mass. : Harvard University Press, 1981.

Moore, Margaret. "Gauthier's Contractarian Morality" in Boucher, David and Paul Kelly, eds., The Social Contract from Hobbes to Rawls, New York: Routledge, 1994.

Nozick, Robert. Anarchy, State, and Utopia, New York: Basic Books, 1974.

Schelling, Thomas. The Strategy of Conflict, Cambridge: Harvard University Press, 1960

Skyrms, Brian. The Stag Hunt and the Evolution of Social Structure, New York: Cambridge University Press, 2004.

Skyrms, Brian. Evolution of the Social Contract, New York: Cambridge University Press, 1996

Sugden, Robert. "Contractarianism and Norms." *Ethics* 100 (1990): 768-86.

Sugden, Robert. The Economics of Rights, Co-operation & Welfare, New York: Blackwell, 1986.

Thomas, Laurence. "Rationality and Affectivity: The Metaphysics of the Moral Self" *Social Philosophy Policy* 5:2 (1988).

Viminitz, Paul. "A Proof that Libertarianism is Either False or Banal" *Journal of Value Inquiry* 34 (2000): 359-367.