Retributive Justice and Social Cooperation

Gerald Gaus

I. The Bishop and John Stuart Mill

In response to American criticisms of the Scottish ministers' decision to release the Lockerbie bomber on compassionate grounds, Cardinal Keith O'Brien, the Roman Catholic Archbishop of St. Andrews and Edinburgh, "launched a scathing attack" on American politicians and, in general, the "culture of vengeance" in the United States. It was reported that he proclaimed that "many Americans were more interested in retribution than justice."[1] For Cardinal O'Brien, retribution and vengeance are closely related—and to be disparaged—and are opposed to an interest in justice. John Stuart Mill offered a rather different analysis; as he saw it, an "essential" ingredient in the "sentiment of justice" is "the desire to punish a person who has done harm."[2] Mill traces the idea of justice back to "authoritative custom," composed of social rules enforced through punishment.[3] "It would always give us pleasure, and chime in with our feelings of *fitness*, that acts which we deem unjust should be punished. . . ."[4] Like the bishop, Mill is worried that the desire for retaliation or vengeance—which, along with sympathy he sees as providing the emotional foundations of justice[5]—is not itself moral and must be directed to moral ends, but he insists that the desire to punish is fundamental to the very idea of justice.

The bishop talks about "retribution," Mill about "punishment."[6] What is their relation? We should distinguish three levels at which "retributivism" is a "theory of punishment." At what I shall call the *shallow* level, retributivism is an account of genuine punishment. As Kurt Baier—very much in the spirit of Mill's analysis—argued, "'punishing' is the name of only a part-activity belonging to a complex procedure."[7] Punishment is part of the complex idea of a system of justice, which is a "whole 'game,' consisting of rule-making, penalization, finding guilty of a breach of a rule, pronouncing sentence, and finally administering punishment."[8] On Baier's analysis, punishment is part of this complex — the idea that hardship is inflicted on a person for an infraction of the rules:

> A method of inflicting hardship on someone cannot be called "punishment" unless at least the following condition is satisfied. It must be the case that when someone is found "not guilty" it is not permissible to go on to pronounce sentence on him and carry it out. For "punishment" is the name of a method, or system, of inflicting hardship, the aim of which is to hurt all and only those who are guilty of an offense. . . . To say that it is of the very nature of punishment to be retributive, is to say that a system of inflicting hardship on

someone could not be properly called "punishment," unless it is the aim of this system to hurt all and only those guilty of an offense.[9]

In this shallow but important sense, there is a large body of evidence that most people *are* "retributivists": they believe that people who do wrong should have hardship inflicted on them just because they have done wrong. They believe that, as Mill said, it is *fitting* that wrongdoers should be punished. Stanley Benn appeared to think that Kant was such a retributivist when he maintains "that the punishment of a crime is right in itself, that is, that it is *fitting* that the guilty should suffer, and that *justice . . . requires the institution of punishment*."[10] Benn charges that this, "however, is not to justify punishment but, rather, to deny that it needs any justification."[11] Benn is right that so understood, this is not a special justification of the practice of punishment; it expresses, rather, the Millian-Baierian observation that the idea of a just system of rules includes punishment, and such punishment is inherently retributive in the sense that Baier indicates above.[12]

Many philosophers find this unsatisfying because they understand "retributivism" in a much stronger sense: as a theory of *why the practice of (retributive) punishment is justified rather than some other practice of rule enforcement*.[13] Retributivism in the shallow sense is a backward-looking doctrine that says that one should be punished if and only if one is guilty of an offense. What are called "retributivist justifications of punishment" are typically doubly backward looking: they hold that this backward-looking practice is justified because of some other backward-looking consideration, such as the importance that people get what they deserve, where desert itself is a backward-looking consideration apart from guilt according to the rules of justice.[14] In contrast, a utilitarian justification of the practice of punishment seeks to give a forward-looking justification of this backward-looking practice of punishment, with all the obvious problems that entails.[15] In this deeper sense, most people are *not* retributivists, for most do not have a justificatory theory of the practice of punishment at all.[16] In between the shallow and this deep understanding is retributivist theory understood as an account of the appropriate severity of punishment: *lex talionis*, the idea that the punishment must have a certain proportionality to the evil of the offense.[17]

In this chapter, I provide an analysis of why the game that Mill and Baier observe, that the practice of justice includes the idea of "retributive punishment"—the idea that the guilty should have hardships inflicted on them just because they are guilty—is one that solves the problem of stable human cooperation in ways that theories of "telishment" or forward-looking enforcement, cannot.[18] In the case of telishment, the institution of enforcing rules via sanctions is not retributive: telishment explicitly aims at a "*telos*," or an end, and so is forward looking in its decision whether harm or costs are to be inflicted as a way to enforce conformity to our rules of social cooperation. The idea that the shallow conception of retributivism is uninteresting because it simply identifies retributivism with punishment overlooks the really interesting question: why do we have a game of justice in which rules are enforced via the backward-looking idea of retributive

punishment rather than some forward-looking strategy? On the face of it, there would seem strong reasons why we should enforce in a forward-looking way—that we should "telish" rather than punish. As Mill stressed, justice is crucial to utility[19]: if the aim of a system of justice is to allow humans to cooperate on mutually beneficial terms, it would seem that when applying sanctions, such a system should have an eye toward promoting cooperation rather than turning our back on the effects of our sanctioning and crying over spilled milk. If we conceive of humans as being devoted to pursuing their own ends and reasonably believe that the game of justice is conducive to this, why are we (or at least most of us) shallow retributivists? Why, *pace* the good bishop, is a culture of justice a culture of retribution?

II. Why Being Nice Is Not Enough

Let us begin by rounding up the usual suspect, in this case the Prisoners' Dilemma in Figure 1. The Prisoners' Dilemma is an example of a family of games, all of which model a conflict of motives. As I have constructed the game in Figure 1, the best payoff is 4, and the worst is 1. On the one hand, we can see the benefits of mutual cooperation: if Alf and Betty cooperate, they both receive a payoff of 3, which is better for both of them than the payoff of 2 that they will receive if they both defect. On the other hand, each has an incentive to defect. If Betty cooperates, Alf receives a payoff of 3 if he cooperates but 4 if he defects, so he does better by defecting. If Betty defects, Alf receives 1 by cooperating but 2 by defecting, so again he does better by defecting. Whatever Betty does, Alf does best by defecting, making defecting his *dominant* strategy: no matter what Betty does, he should defect. And Betty's reasoning is *mutatis mutandis*, the same. But this leads to mutual defection and, as we have seen, they can do better by mutual cooperation.

		Betty	
		Cooperate	*Defect*
Alf	*Cooperate*	3 3	4 1
	Defect	4 1	2 2

Figure 1: A Prisoners' Dilemma in Ordinal Form

More generally, we can see how social interactions with this structure are an impediment to mutual advantage. If we suppose that individuals are solely devoted to their own ends, the development of mutually advantageous market

exchange—which depends on trust—seems difficult to explain. As Hobbes so effectively showed, individuals who are solely devoted to their private ends will be sorely tempted to renege on "covenants": if the other party performs first and so gives the second party what she wants, there seems to be no incentive for the second party to perform her part of the bargain.[20] Rather than exchanging, she will be tempted to snatch the goods and flee.[21] Given sufficiently narrow, self-interested utility functions, she will often be tempted to snatch—getting the good without paying for it.

Clearly, the development of a system of rules of justice that requires each to cooperate would be beneficial to each, but, so long as they reason simply in terms of their own interests, they will be tempted to violate the rules. For Hobbes, once we recognize this, we see why a system of punishment is necessary to secure justice: there must be an authorized punisher who can inflict such hardships on defectors that once they tally up the gains from defection but subtract the costs of the imposed hardships, they will see that their interests align with cooperation.

As we will see, this is indeed a fundamental insight and one of the reasons anyone who grapples with the problem of cooperation in such "mixed motive" interactions must come to terms with Hobbes. But we might think that Hobbes's conclusions about the necessity of enforcement only follow because the Prisoners' Dilemma supposes that people are pretty nasty. According to Prisoners' Dilemma preferences, given a chance to cheat on Betty when she has already done her part of the bargain, Alf prefers to take advantage of her rather than do his part. We can imagine the bishop saying that it is precisely such an individualist (American?) culture that breeds the culture of vengeance. But people are nicer than that. The people modeled in the Prisoners' Dilemma game order the options like this: (1) the other cooperates and I cheat; (2) we both cooperate; (3) we both cheat; (4) I cooperate, and the other cheats. But nice people would not wish to cheat; they would prefer to cooperate if the others do so.

		Betty	
		Cooperate	*Defect*
Alf	*Cooperate*	4 4	3 0
	Defect	0 3	3 3

Figure 2: A Stag Hunt

So suppose we reorder the outcomes so that the best option for both is mutual cooperation. But neither should we suppose that Alf and Betty are saintly: they still think the worst outcome is to be played for a sucker, cooperating while the other defects. This gives us a "Stag Hunt" as in Figure 2. In this game, rational individuals are pulled in two directions: mutual benefit points to cooperating (each gets 4), but risk aversion points to defecting (a guarantee of getting 3 is better

than a chance of getting 0). There are two equilibria in this game: mutual cooperation and mutual defection. Under some conditions, it is remarkably easy to end up in social states in which everyone defects. Suppose first, a very simple iterated (repeated) game in which individuals randomly encounter each other, playing a series of games, and suppose (now using cardinal payoffs) that defecting has a payoff of 3 and mutual cooperation 4, but cooperating alone gives a 0 payoff. If less than 75 percent of the population begins with cooperating, the population will gravitate over time to all defectors; only if the initial cooperating population is over 75 percent of the total will the group evolve into all cooperators.[22]

So it is not only important that lots of people have "nice" Stag Hunt outlooks to assure that they will abide by the rules of cooperation. Alf must be confident that Betty is nice too, and, moreover, he must know that she knows he is nice (and that he knows she knows this, and so on). If they are cautious and start out by playing "defect," even a population of all "nice" people can end up *all* defecting. If the conversion process is incomplete—if most but not all of us have adopted "nice" preferences such as those in Figure 2, but some remain unreformed, Prisoners' Dilemma players—and this fact is generally known—the prospects for social cooperation seem quite dim. Gregory Kavka introduced the game of an "Assurance Dilemma" to model interactions in a Hobbesian state of nature, according to which some agents have nasty Prisoners' Dilemma orderings while others have nicer, more cooperative orderings (as in Figure 3).[23] Assume that parties do not have perfect information about each other's preferences. When Alf encounters another, the question facing nice Alf is whether he is playing another nice person—in which case he essentially will be playing the game in Figure 2—or whether he is playing a nasty person, as in Figure 3. Here, even a nice Alf should defect, for his cooperative move will be met with defection from Betty, giving him his worst outcome. Note that the sole equilibrium in this game is the same as that in the Prisoners' Dilemma: mutual defection.

		Betty (Nasty)	
		Cooperate	*Defect*
Alf (Nice)	*Cooperate*	3 4	4 1
	Defect	1 2	2 2

Figure 3: An Assurance Dilemma

Peter Vanderschraaf has recently developed a far more sophisticated version of this game in which players have a range of possible payoffs (rather than just two types), and ways of estimating (based on past performance) what the payoffs of the other player are. Players are uncertain about the payoffs of others but have ways of learning and updating their probability estimate that the others are prone

to attack. In computer simulations of anarchy under these conditions, Vanderschraaf finds that "[t]he presence of only a few 'nasty' individuals gradually drives all, including those inclined to be 'nicer,' to imitate the 'nasty' conduct of these few. This dynamic analysis suggests that the Hobbesian war in anarchy is indeed inevitable in most realistic circumstances."[24] Under conditions of imperfect information and a range of parties' preferences, the uncertainty created by the presence of a few nasty agents (that is, with Prisoners' Dilemma-like preferences) leads to universal war (or defection).

Vanderschraaf's analysis is important as it illustrates a deep flaw in a long line of theorizing that has supposed that social cooperation emerges from people reasoning themselves into, or adopting, more cooperative preferences. It has often been thought that the core task in explaining the rise of cooperation is to show the reasonability of preferences for conditional cooperation ("I'll cooperate if you do"). That is, at best, only the first step. Having a large majority of the population with preferences for conditional cooperation (such as in Figure 2) by no means guarantees social cooperation or even makes it likely. Although we began with the Prisoners' Dilemma, with "nasty" preference orderings, we now see that the problem is not only the character of the preferences but also uncertainty about what others' preferences are. Even interactions that are "objectively" cooperative—that is, under perfect common knowledge, we would all see that we are conditional cooperators—may easily end up at noncooperative equilibria. Indeed, under full common knowledge and cooperative preferences, cooperative equilibria are by no means certain; as Cristina Bicchieri stresses, firm first-order expectations about what others will do are required.[25] It turns out that the problem of social cooperation is difficult for even a population of cooperative people.

III. Minimal Enforcement: Boycotts

In addition to being nice, perhaps it will suffice if people are also discriminating. If cooperators can distinguish other cooperators from defectors, and interact with the former and boycott the latter, they could gain the benefits of cooperation without opening themselves up to be suckers. More formally, we can say that when cooperation is correlated—when there is a tendency for cooperators to seek out, identify, and interact with other cooperators—cooperative interaction can flourish.[26] We can think of this as a strategy of boycotting defectors, a relatively mild form of sanctioning for noncompliance with cooperative norms. Boycotting is effective *if* a cooperator has sufficient information about the dispositions of others—and, of course, that's the rub. Defectors have a good incentive to appear cooperative, so without a good deal of widespread, accurate, and shared information about who is a cooperator and who is a defector, boycotting will be an inefficient mechanism to produce compliance.

Hume and Hobbes thought that one's *reputation* as either a cooperator or defector—the public knowledge of one's past behavior—would serve as an

effective way to enforce contracts. Again, Vanderschraaf's work has enlightened us about the conditions for compliance with covenants through reputation. Vanderschraaf focuses on what he calls "the covenant game," seeking to model Hobbes's and Hume's proposal that rational agents would not cooperate with those who have reputations as defectors, and this knowledge should itself provide a sufficient incentive for would-be defectors to refrain from double-crossing on covenants.[27] In his iterated (repeated) "covenant game," one has a choice to (1) boycott the other party by refusing to make an agreement; (2) promising and performing; or (3) promising and then double-crossing; as in the Prisoners' Dilemma (which forms a subgame of Vanderschraff's more complex game) there is an incentive to double-cross rather than perform. The parties know that they would gain in the present stage by double-crossing but also may gain a bad reputation by double-crossing, and those with bad reputations would be boycotted in future stages of the game and so would not gain in those stages the fruits of social cooperation.

The novel and insightful idea in Vanderschraff's analysis is modeling types of information available to the parties. If we possess accurate common knowledge of the trustworthiness of potential partners—we all know about who is trustworthy and all know that we all know this—reputation effects can indeed secure cooperation. But public, accurate, common knowledge is seldom available. Vanderschraaf thus models information produced by gossip: individuals letting other individuals know what they think of certain parties, taking account of the effects of false gossip as well as true gossip. As ethnographic studies of cooperation via reputation have shown, there are strong incentives to exaggerate and fake one's cooperative history.[28] Once we model information as gossip (true and false) and allow defectors to adopt slightly more sophisticated strategies (such as only double-crossing half the time rather than always), Vanderschraaf shows that such defectors "can fare better than Humeans in a community that must rely upon private information or 'gossip' only to spread information."[29]

IV. Getting Tougher: Deterrence

Perhaps it is time to get a little tougher on defectors. Robert Axelrod is widely understood to have shown that cooperation can emerge among "egoists" (people with Prisoners' Dilemma-type orderings) playing repeated Prisoners' Dilemmas.[30] To see this, assume that although you are in a Prisoners' Dilemma with Alf, both of you know that every day from here on in, you will play one Prisoners' Dilemma after another. Axelrod has shown, using a computer simulation, that you often would do best by adopting a very simple strategy: tit for tat. According to tit for tat, your first move is the cooperative one. But if Alf defects rather than cooperates, the next time you meet Alf, you will be uncooperative, too. In short, except for the first move of the game, you decide whether to cooperate or act aggressively with respect to any person by a simple rule: "I'll do to him on this

move whatever he did to me the *last* time we met." Essentially, a tit-for-tat player says to others, "If you defect on me in this game, you will get away with it, but I guarantee that in the next game, I will defect so that we will both be losers. But I am forgiving: if you start cooperating again, I'll begin cooperating again on the move after your cooperative move." Axelrod constructed a computer program specifying a game with hundreds of moves and a variety of actors employing different strategies. Some always act cooperatively no matter what others do, while some always defect. Each strategy played a 200-move game with each other strategy, making for over 100,000 moves: tit for tat won every time.

Many believe that tit for tat is somehow the correct strategy for all repeated Prisoners' Dilemmas, but this is not so.[31] Whether tit for tat is the correct strategy depends on the strategies of the other players. There is no single solution to an infinite (or indefinite) series of Prisoners' Dilemmas; indeed, repeated Prisoners' Dilemmas have infinitely many equilibrium "solutions." Two tit for tatters are indeed in a Nash equilibrium—neither will benefit from a unilateral change of move—when cooperating in such a series. Recall that a tit for tatter cooperates on the first game in a series of Prisoners' Dilemmas and then will do to its opponent on game n whatever its opponent did to it on game n-1. So if one tit for tatter defects on game n-1, the other tit for tatter responds by "punishing" the defector in game n. If one tit for tatter unilaterally defects from cooperation, the other tit for tatter will also "punish" and so lower the payoffs of the defector. Knowing this, neither tit for tatter can gain by unilateral defection, so they are in a Nash equilibrium.

But it is not just two tit for tatters that are in equilibrium—consider "the Grim strategy." Grim cooperates on the first move, but if its opponent defects, Grim will "punish" for every move after that—forever. Two Grim players are also in equilibrium: neither would benefit from defection. The important thing here is that what are generally called "punishing strategies" can achieve equilibrium: if I can inflict hardship on you I can deter you from defecting in cooperative interactions. Repeated Prisoners' Dilemmas allow what are essentially "self-policing contracts" to cooperate. Since we are playing infinitely or indefinitely many games, I can afford to inflict hardships on you now to bring you around, and, seeing that, you will not unilaterally defect. Indeed, any cooperative outcome that gives each player more (or, more formally, at least as much) as the minimum he might receive if the other player acted in the most "punishing" way can be an equilibrium: if we are each above this minimum point, then one party still has room to inflict hardship on the other for a deviation from the "contract." This is the baseline, the payoff a person could get if her opponent was intent on making sure she got as little as possible. As long as the agreement (the coordinated payoffs) is above this minimum, there is room for "punishment," and so unilateral defection will not pay, and thus there will be a Nash equilibrium. This result is known as "the folk theorem."

I have referred to these as "punishing" strategies since that is the common term in the literature, but we can see that they are not engaging in retributive punishment—such strategies adopt a form of deterrence. They make the following threat: if you defect at time t_1, I will inflict costs at a later time t_2 *as a way of securing future cooperative behavior.* As rational individuals, each player is solely devoted to maximizing her payoffs; but given that, "punishment" is simply an investment *now* so as to secure better behavior *later*. In these sorts of iterated interactions, a player is taking account of two factors: (1) how well he can do in this game, and (2) whether he can deter the other from defecting in the future. The second consideration is sometimes called the "shadow of the future." The longer it is—the further the future extends—the more people engage in deterrence and cooperative behavior. However, as the shadow of the future shortens—as people come to believe that they are coming near to the end of their interactions, and so there is less opportunity to be "punished" and less likelihood that one's inflicting "punishment" on them will affect their behavior, people switch from cooperative to defecting behavior.[32] Given this, such deterrence is effective in securing cooperation only when the deterring person has expectations for extended future interactions. If I am a defector, and I know we will never play any more games, I know that you will not incur additional costs to punish me for defecting in this interaction because it would just make you even worse off. As social organization expands, we are constantly encountering people whom we have never met before and whom we are uncertain whether we will meet often again, and so very often, it will not be rational to invest in deterrence behavior. Sophisticated models indicate that deterrence (like tit for tat) allows the development of cooperation in small groups but not in larger ones. [33]

V. Retaliation: Its Virtues and Puzzles

As we just described, cooperation among tit for tatters (or people playing related strategies) unravels as the end of the game approaches; the less certain we are of future interactions, the less incentive we have to sanction noncompliers, and so the less incentive for them to comply. The clear problem is the resolutely forward-looking nature of such telishing strategies: only if the expected benefits of the sanction exceed its costs will rational tit for tatters telish. Would a "culture of vengeance" help?

Call tit for tatters who are resolutely backward looking—who punish rather than telish—*vengeful*: they engage in true punishment because they impose sanctions at t_2 *just because* of the defection of another at t_1, regardless of the shadow of the future. Consider the famous case of private enforcement in medieval Iceland.[34] The rules of social cooperation were enforced by the victims of defection: injured individuals had the right to extract fines (socially fixed at different levels for different offenses) from those who had harmed them through a violation of the rules. A victim could either pursue the fines himself or transfer (by contract) the

right to exact them to a third party. The system sought to reduce the costs of self-enforcement; for instance, society as a whole did become involved if the offender refused to pay, declaring the offender an outlaw, which affected what others could do. But the offender might resist; if he did so, any harm done in wrongful resistance was considered yet another offense, which could be the grounds for yet another self-enforcement action, and so on.

Anarchists tend to be fans of the Icelandic system, and they insist that the records do not show that it led to escalating feuds (as vengeance apparently did in the Hatfield–McCoy feud of 1878–1891). The key feature of the system is that the total costs of enforcement were put onto the victim. Suppose the costs of the violation to the victim are c, the costs of enforcement to the victim are d, the fine is f, and the probability of successfully exacting the fine is p.[35] So, unless the expected fine pf equals or exceeds the total costs $c + d$, the victim loses out. The problem is that if p is low (the fine is unlikely to be exacted), and if d is high (and increases in d do not sufficiently raise the possibility of successful enforcement), then unless the fines are set very high (which, one would think, would increase the tendency of defectors to resist, which would itself lower p), the victim will be faced with a double loss: she already has lost c due to the offense, and now she will incur an additional loss of d on the enforcement.

It is reported that the Icelandic system sought to increase p by only including crimes with a probability of detection near unity, and the threat of being declared an outlaw would reduce the probability of stiff resistance.[36] Nevertheless, even an advocate of the system such as David Friedman allows that in some cases it may have been that $pf < c + d$, and so the victim would have lost out.[37] Friedman argues that enforcement would nevertheless occur because one wished to have a reputation as an enforcer, but we have already seen that reputation is an unreliable mechanism, at least in larger groups (which, admittedly, medieval Iceland was not). Here is the crux of the problem: economists such as Friedman wish to show that victims generally acted as good rational economics agents, only engaging in retaliation if $pf \geq c + d$, but it is very hard to think an enforcer will be confident that this will be so, at least in large societies where defectors may resist.

To stabilize cooperation in the face of defection, it would seem two things are required. First, enforcers must have a strong incentive to engage in retaliatory action even when the gains from retaliation are less than the costs. If enforcers were not so calculating and rational, they would retaliate even when it was a net loss for them to do so; the probability of enforcement would then clearly go up. Thus, if instead of being a rational policy, retaliation was a basic emotion or desire—if we enjoy retaliation—the certainty of enforcement would rise. But given that all the weight of enforcement is put on the shoulders of the victims (or their designated agents), it might seem that they would have to *really* enjoy their lone enforcement activity. To use Shaun Nichols's quip, they would have to think that retribution was very *yummy* indeed![38] It would help if the desire to retaliate was not just personal, so third parties would share some of the costs with the victims.

This is the second of the two sentiments that Mill thought was required for punishment: sympathy for the wronged and a desire to strike back on their behalf.[39]

To show that these two sentiments of punishment would stabilize cooperation in larger groups is not to say that there is any reason to believe that these sentiments would arise. Just because they would be useful does not mean that we would acquire them. Moreover, their acquisition seems especially perplexing. Up to now, we have been considering enforcement mechanisms that are part of a rational system of cooperation—a system in which people are adding up the costs and benefits and deciding to cooperate and enforce. We have been investigating the limits of calculating instrumentalist enforcement. But through the last sections, we have been led to the conclusion that the more we think of enforcement in terms as an investment to obtain the future goods, the less certain we are of securing it. The enforcement of cooperative norms is a great public benefit to the group as a whole, but it often does make sense for an individual to incur the costs of enforcement herself. What we need, it seems, is an account of how it can be good in terms of cooperative order for people not to think about enforcement in these calculating terms.

VI. A Simple Evolutionary Model

Instead of supposing that we are confronted with rational agents calculating the most efficient ways to achieve their ends, suppose instead we take an evolutionary perspective. Suppose, that is, we think of a society as populated by fixed strategies: some always cooperate, some always defect, and so on. In this sort of model, agents do not change their moves (as I said, they are fixed). Rather, the model is based on *replicator dynamics* in which those strategies that tend to have higher average payoffs increase their share of the population and so displace lower payoff strategies.[40] So what "moves" is the percentage of the population employing a certain strategy, as some replicate faster than others.

Let us begin with a very simple model, according to which these people are all Unconditional Rule Followers: they are unconditionally disposed to follow rules that dictate cooperation in mixed motive games such as the Prisoners' Dilemma, regardless of the strategy of other actors. Now a group with a preponderance of Unconditional Rule Followers will outperform groups of purely instrumental agents: as we have seen, instrumentally rational agents have considerable difficulty stabilizing cooperative interactions. In Prisoners' Dilemmas, a population of Unconditional Rule Followers will always receive their second highest payoffs. Thus, if we consider things from the level of group competition, social groups composed of Unconditional Rule Followers will do much better than groups of instrumentally rational agents. The people in them will have more of their ends satisfied, and they will generally achieve higher standards of living. In competition with other groups, we would expect them to displace less efficient

groups.[41] Perhaps more importantly, we can expect their forms of social organization to be copied by others groups, as useful norms tend to spread.[42] It must be stressed that this need not be a story about genetic fitness and natural selection. There is considerable evidence that copying the activities of the more successful members of one's group is fundamental to culture.

The problem is that Alf's being an Unconditional Rule Follower easily detracts from his overall fitness within the group, at least if we think that his fitness is well correlated with his achieving his own individual goals. Although individuals in groups dominated by Unconditional Rule Followers will, on average, do better than individuals in groups where, for example, everyone is a Defector, groups of Unconditional Rule Followers could always be invaded by simple Defectors (who always defect on cooperative arrangements), each of whom would do better than the group average, since they would receive the group benefits of social cooperation as well as the individual advantages of cheating in all their personal interactions. But this means that despite their intergroup strengths, groups of Unconditional Rule Followers are not stable, as they are liable to be invaded by nasty types.

This appears to reaffirm from a different perspective our general lesson thus far: how difficult it can be to secure cooperation in the face of defectors! Let us introduce a slightly more complex model. Following Boyd et al., suppose that the benefit to the group of each act of cooperation through following norms is b, and a contributor incurs a cost of c to produce these benefits.[43] Defectors in the group produce no benefits and incur no costs, though by living in a cooperative group, they enjoy the group benefits of cooperation. To simplify matters, assume that all the benefits of each act of cooperation accrue to the entire group. Whatever a person gains by her own b (her contribution to the group's welfare) is less than c (its cost); otherwise there would be no temptation to defect. If there are x Unconditional Rule Followers in the group (where x is less than the total population of the group), the expected benefits to an Unconditional Rule Follower is $bx - c$: the total benefits of x cooperative acts minus the cost of her own cooperation. The expected benefit to Defectors is simply bx, and so the Defector will always have a higher expected payoff than the average in any group of Unconditional Rule Followers. So far it seems that while groups with large numbers of Unconditional Rule Followers will thrive, within those groups, they will always be less fit—do less well—than Defectors. Given this, Defectors would provide a model of someone "doing better," and so imitators would tend to copy them, spreading defection throughout the group.

Now suppose that "Rule-Following Punishers" enter the population. As the name indicates, Rule-Following Punishers unconditionally follow cooperative rules and *always* punish those who do not (in the spirit of Kant, even if their society were about to be dissolved, they would still punish the last Defector). Again following Boyd et al., we can say that a punisher reduces the benefit of an

act of defection by punishment p at a cost of k to the punisher. Suppose that there are y punishers who punish each act of defection. It follows that:

1. The expected benefits to a nonpunishing Easygoing (unconditional) Rule Follower (an x-type person) are: $b(x + y) - c$. Each Easygoing Rule Follower gets the full advantages of cooperative acts performed by both her own x-types and well as the Rule-Following Punishers (y-types) and, in addition, she receives the social advantages of punishment of Defectors by Rule-Following Punishers and incurs only the cost of her own cooperative act.

2. The expected benefits to a Defector are $b(x + y) - yp$. A Defector still receives the full benefits of cooperation generated by x and y types (with no cost of cooperation) but incurs a punishment of p from each of the y punishers.

3. The expected benefits of being a Rule-Following Punisher (a y-type) are: $b(x+ y) - c - [k(1 - x-y)]$. Each y receives the full benefits of cooperation generated by x and y types, and like Easygoing Rule Followers, each incurs the costs of cooperation. In addition, each punisher must incur a cost of k for punishing each Defector; if there are x Easygoing Rule Followers, and y Rule-following Punishers in a group, then the number of Defectors (normalizing the population to one) is $1 - x-y$.

Therefore, if the costs of being punished are greater than the costs incurred in cooperating—that is, if $yp > c$—Defectors will decline in the population.

So this solves our long-standing problem: groups with Rule-Following Punishers are not liable to invasion by nasty types. Have we shown how social cooperation can be stabilized? Not quite. Unfortunately, they are liable to invasion by nicer types! The problem is that Rule-Following Punishers have a fitness disadvantage relative to Easygoing Rule Followers. Groups of Rule-Following Punishers can be invaded by Easygoing Rule Followers, who reap all the benefits of social cooperation and the elimination of nasty types but do not incur the costs of inflicting punishment. On the face of it, just as the superior fitness of Defectors within the group undermined Generalized Rule Followers, the superior fitness of Easygoing Rule Followers undermines the Rule-Following Punishers. However, as Boyd et al. show, the crucial difference between the two cases is that in the first, since the Generalized Rule Followers incur a constant cost c, they are always less fit by c in relation to Defectors—they can never, as it were, close the gap. However, the gap in fitness between Easygoing Rule Followers and Rule-Following Punishers reduces to zero if Defectors are eliminated from the population; it is only the punishing of Defectors that renders Rule-Following Punishers less fit than Easygoing Rule Followers.

We would not expect the fitness gap to reduce to zero—mutations, errors in determining defection, and immigration from other groups will result in above zero rates of punishing. Nevertheless, under a large range of values of the relevant variables, the discrepancy in relative fitness is small enough so that Rule-

Following Punishers remain a high proportion of the population, and defection thus declines. Moreover, cooperation on the basis of Rule-Following Punishers is stabilized if we introduce *n*-level punishments by which a person is punished at the level *n* for failing to punish at level *n*-1. Now the Easygoing Rule Followers are also subject to punishment for their failure to punish. It can be shown that, those who punish at the n^{th} level are always less fit that those who do not punish *at that level*, the difference in fitness is diminishing at each level. At some point, the difference in fitness is sufficiently small that a slight tendency to conform to the majority retributivist practice can switch everyone over to being a punisher.[44]

It is important that in such group selection models, the survival of a strategy depends on two factors: (1) its individual fitness within the group, and (2) its inclusion in a group that has group-wide advantages over competing groups. In pure individual fitness accounts, of course, only the first factor is relevant. In our first group selection account with only Generalized Rule Followers and Defectors, the former enhance group fitness, but their serious individual deficit drives them to extinction within the group, and the result is that Defectors take over. Interestingly, although Defectors take over and, indeed, a group of all Defectors cannot be invaded by Generalized Rule Followers, the group would be driven to extinction in a competition with more cooperative groups. In the case of a mixed population of Easygoing Rule Followers, Defectors, and Rule-Following Punishers, the Rule-Following Punishers reduce the number of Defectors; during an initial period, they are less fit than Easygoing Rule Followers, but as Punishers succeed in eliminating Defectors, their fitness converges with their more easygoing comrades, and the group's overall fitness is enhanced. The more the Rule-Following Punishers must punish (say, because of high rates of mutation that constantly reintroduce Defectors into the population), the greater will be the gap between the fitness of the Punishers and the Easygoing Rule Followers.

This discussion has been rather complex, but the core point is straightforward. The evolutionary model we have been examining shows how individuals who possess Mill's two sentiments of punishment—a desire to retaliate on the basis of past wrongs and sympathy with wronged others, resulting in a desire to seek revenge against wrongdoers—are more apt to form social groups in which the population's goals are satisfied than are groups of agents who only see enforcement in instrumental terms and which are resistant to invasion by nasty types. The groups populated by such individuals do much better than groups without them (they are bad news for Defectors), and within these groups, the Rule-Following Punishers may be nearly as fit as the cooperative free riders, and they also keep Defectors from invading the group. Plus, they can pretty much eliminate Easygoing Rule Followers by punishing those who fail to punish Defectors. To be a member of a society dominated by Rule-Following Punishers is the most effective way to advance one's ends.

VII. The Critic, the Retributivist, and the Evolution of Social Order

Our punishers are genuinely "altruistic" agents in the sense that they engage in activities that do not best maximize their own goals but are advantageous for the group. Purely instrumental agents could never reason themselves into being strictly altruistic—performing acts that, all things considered, set back their own values, goals, and ends.[45] We thus have a powerful model for one of the most perplexing of evolutionary phenomena: the selection of altruistic traits which, by definition, reduce the individual fitness of those who possess them. Punishing violators is a public good; because the Easygoing Rule Followers share the benefits of such punishment but do not incur the costs of contributing, they do better. And, of course, an agent solely devoted to her own goals will defect if she is not punished; Rule-Following Punishers will not. Thus, if we focus on simply instrumental rationality in terms of promoting one's goals narrowly understood (that is, apart from a taste for retribution), Rule-Following Punishers are not instrumentally rational.

Under some configurations of payoffs, models of the evolution of social cooperation show a mixed equilibrium of populations split between punishing and easygoing cooperative strategies (perhaps with some Defectors as well).[46] I began by noting the conflicting views of the bishop and John Stuart Mill: for one retaliation is distinct from, and often opposed to, justice; and for the other, retributive punishment and its accompanying sentiments are part and parcel of the very idea of justice. It has long been the case that people are divided in this way. Evidence indicates a majority with retributivist views about justice, almost always with a spirited dissent by some. The interesting possibility arises that *this* is our evolutionary stable outcome: a mixed population whose members have different views of the relation of justice and retribution.

It is easy for intellectuals to dismiss the shallow retributivism of the *hoi polloi*. It is not based on the deep philosophical justifications to which intellectuals are so drawn and which they love to construct (and destroy); it is a competency about the "game" of justice on which our cooperative order is based. If asked, "why *this* conception of a just order?" rather than one based on boycotts, deterrence, or Icelandic private enforcement, the *hoi polloi* will be unable to answer. Like so much of our social world, this system was not the product of philosophic construction but of social and biological coevolution: it is a crucial element in solving the absolutely fundamental problem of stabilizing social cooperation. If we did not have a "culture of vengeance," the good bishop probably would not be in the position to insist on compassion (a society of easygoing types can be invaded easily by nasty types). Yet, as Mill so clearly recognized, shallow retributivism needs to be moralized. We have a taste for punishment and, like all tastes, we may indulge it excessively. We need critics like the bishop to remind us that punishment imposes hardships for what is already done and in the past. A just and humane society seeks to stay fit and trim in satisfying such tastes.[47]

Notes

[1] *Financial Times* (U.S. edition), Monday August 9, 2010, p. 2 (News Digest).

[2] Mill, *Utilitarianism*, ch. V, ¶18.

[3] Ibid., ch . V, ¶12.

[4] Ibid., ch. V, ¶13 (emphasis added).

[5] Ibid, ch. V, ¶20.

[6] Mill also refers to "retribution" twice in chapter V of *Utilitarianism*. In paragraph 29, he uses "retribution" to describe a particular principle of justice—*lex talionis*—concerning how *much* punishment is appropriate (so not a general justification of punishment itself); in paragraph 34, it is linked more generally to the sentiment of vengeance.

[7] Baier, "Is Punishment Retributive?"

[8] Ibid., p.26.

[9] Ibid., p. 27.

[10] Benn, "Punishment," p. 30 (emphasis added).

[11] Ibid.

[12] Benn concurs, I think, that retributivism can be understood in such a way that it is a definition of punishment. See Benn and Peters, *Social Principles and the Democratic State*, p. 184.

[13] See ibid., p. 175ff.

[14] If it is not distinct, it collapses into the shallow idea that justice requires that the guilty be punished. I defended a version of this deeper understanding of retributivism in "Taking the Bad with the Good." For an extended justification of deep retributivism, see Moore, "Moral Worth of Retribution."

[15] For what is still a good account of these, see Benn, "Punishment."

[16] As Nichols shows in "Brute Retributivism."

[17] See Benn, "Punishment," p. 32, and Mill's view in note 6 above. Moore insists that retributivism is *only* a deep theory, and so denies that this "*lex talionis*" sense, or the shallow sense, are retributivist theses at all ("Moral Worth of Retibutivism," pp. 188–89).

[18] This term derives from Rawls, "Two Concepts of Rules," p. 27.

[19] Mill, *Utilitarianism*, ch. ¶37.

[20] Hobbes actually thinks a person has some reason to perform second, but this is usually too weak to outweigh her selfish passions (see *Leviathan*, ch. 14).

[21] On the game of snatch, see Schwab and Ostrom, "Vital Role of Norms and Rules," p. 205ff.

[22] See Skyrms, *Stag Hunt and the Evolution of the Social Structure*, ch. 3.

[23] The "Assurance Games" is a slight variation on the Stag Hunt in Figure 2. While in the Stag Hunt, I am indifferent between "both of us defecting" and "me defecting while the other cooperates," in the Assurance Game I prefer "I defect while the other cooperates" to "both of us defecting." The difference is not important here. Vanderschraaf ("War or Peace?") reports this is Kavka's view on

the basis of the latter's unpublished 1989 manuscript on "Political Contractarianism."

[24] Vanderschraaf, "War or Peace?," p. 245.

[25] Bicchieri, *Grammar of Society*, p. 37.

[26] See Skyrms, *Evolution of the Social Contract*, ch. 3.

[27] Vanderschraaf, "Covenants and Reputations."

[28] Henrich and Henrich, *Why Humans Cooperate*, ch. 6.

[29] Vanderschraaf, "Covenants and Reputations," p. 184.

[30] See Axelrod, *Evolution of Cooperation*.

[31] See Binmore, *Natural Justice*, p. 78ff.

[32] This is a well-established finding; see, for example, Selten and Stocker, "End Behavior in Sequences of Finite Prisoner's Dilemma Supergames."

[33] See Richerson and Boyd, *Not by Genes Alone*, pp. 199–201. For an analysis showing the difficulty for reciprocity developing when larger groups face iterated N-person Prisoners' Dilemmas, see Boyd and Richerson, *Origin and Evolution of Culture*, ch. 8.

[34] I am following here Friedman, "Private Creation and Enforcement of Law."

[35] The variables p and d are not independent; one can increase p, the probability of success, by increasing d, say by purchasing weapons.

[36] Friedman, "Private Creation and Enforcement of Law," p. 592.

[37] Ibid.

[38] Nichols, "Brute Retributivism." Nichols cites brain imaging evidence that people enjoy punishing norm violators.

[39] Mill, *Utilitarianism,* ch. V, ¶19.

[40] See Skyrms, *Evolution of the Social Contract*.

[41] A striking example of how more efficient forms of social organization can lead one group to displace another is that of the Dinka and the Nuer. For an explicitly evolutionary account, see Richerson and Boyd, *Not by Genes Alone*, pp. 23–5. For an empirical study of how group functional norms can evolve by cultural group selection see Boyd and Richerson, *Origin and Evolution of Culture*, ch. 11.

[42] See Boyd and Richerson, *Origin and Evolution of Culture,* ch. 12. Both group displacement and copying the norms of more successful groups are important to Hayek's account of social evolution; see my "Evolution of Society and Mind," pp. 238–46. After a long period of disfavor, group—or more generally, multilevel—selection is again respectable in genetic evolutionary theory. For a helpful analysis, see Okasha, *Evolution and Levels of Selection*. It is more than respectable in accounts of cultural evolution.

[43] Boyd et al., "Evolution of Altruistic Punishment," 215–27.

[44] Boyd and Richerson, *Origin and Evolution of Cultures*, ch. 10.

[45] Sober and Wilson strenuously argue for group selection accounts of altruistic behavior in their *Unto Others: The Evolution and Psychology of Unselfish Behavior*.

[46] See Boyd et al., "Evolution of Altruistic Punishment"; Skyrms, *Evolution of the Social Contract*, ch. 2

[47] I greatly benefited from a long conversation one hot Tucson summer day over cold beers with Shaun Nichols. I also have learned a great deal from graduate students in my seminar in moral and social evolution at the University of Arizona—special thanks to Keith Hankins, John Thrasher, and Kevin Vallier. Some of the material in this chapter is drawn from my forthcoming book, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World* (Cambridge University Press), chapters 3 and 4. My special thanks to Mark White for helping me avoid some errors.

References

Axelrod, Robert. *The Evolution of Cooperation*. New York: Basic Books, 1974.

Baier, Kurt. "Is Punishment Retributive?" *Analysis* 16 (1955): 25–32.

Benn, Stanley I. "Punishment." In *The Encyclopedia of Philosophy*, vol. 7, edited by Paul Edwards, 29–36. New York: Macmillan and the Free Press, 1967.

——and R.S. Peters. *Social Principles and the Democratic State*. London: Allen and Unwin, 1959.

Bicchieri, Cristina. *The Grammar of Society*. Cambridge: Cambridge University Press, 2006.

Binmore, Ken. *Natural Justice*. Oxford: Oxford University Press, 2005.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. "The Evolution of Altruistic Punishment." In *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, edited by Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr, 215–28. Cambridge, MA: The MIT Press, 2005.

Boyd, Robert, and Peter J. Richerson. *The Origin and Evolution of Culture*. Oxford: Oxford University Press, 2005.

Friedman, David. "Private Creation and Enforcement of Law: A Historical Case." In *Anarchy and the Law*, edited by Edward P. Stringham, 586–601. New Brunswick, NJ: Transaction Publishers, 2007.

Gaus, Gerald. "The Evolution of Society and Mind: Hayek's System of Ideas." In *The Cambridge Companion to Hayek*, edited by Ed Feser, 232–58. Cambridge: Cambridge University Press, 2006.

"Taking the Bad with the Good: Some Misplaced Worries about Pure Retribution." In *Legal and Political Philosophy*, edited by Enrique Villanveua, 339–62. Amsterdam: Rodopi, 2002.

Gintis, Herbert. *Game Theory Evolving*. Princeton: Princeton University Press, 2000.

Henrich, Natalie, and Joseph Henrich. *Why Humans Cooperate: A Cultural and Evolutionary Explanation.* Oxford: Oxford University Press, 2007.

Hobbes, Thomas. *Leviathan.* Edited by Edwin Curley. Indianapolis, IN: Hackett, 1994.

Mill, John Stuart. *Utilitarianism*. In *The Collected Works of John Stuart Mill*, edited by John M. Robson. Toronto: University of Toronto Press, 1985.

Moore, Michael S. "The Moral Worth of Retribution." In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, edited by Ferdinand Schoeman, 179–220. Cambridge: Cambridge University Press, 1987.

Nichols, Shaun. "Brute Retributivism." Working paper.

Okasha, Samir. *Evolution and Levels of Selection*. Oxford: Clarendon Press, 2006.

Rawls, John. "Two Concepts of Rules." In *Collected Papers of John Rawls*, edited by Samuel Freeman, 20–46. Cambridge, MA: Harvard University Press, 1999.

Richerson, Peter J., and Robert Boyd. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press, 2005.

Schwab, David, and Elinor Ostrom. "The Vital Role of Norms and Rules in Maintaining Open Public and Private Economies." In Moral Markets: The Critical Role of Values in the Economy, edited by Paul Zak, 204–27. Princeton: Princeton University Press, 2008.

Selten, Reinhard, and Rolf Stocker. "End Behavior in Sequences of Finite Prisoner's Dilemma Supergames." *Journal of Economic Behavior and Organization* 7 (1986): 47–70.

Skyrms, Brian. *The Evolution of the Social Contract*. Cambridge: Cambridge University Press, 1996.

——. *The Stag Hunt and the Evolution of the Social Structure*. Cambridge: Cambridge University Press, 2005.

Smith, John Maynard. "The Evolution of Behavior." *Scientific American* 239 (1978): 176–92.

Sober, Elliot, and David Sloan Wilson. Cambridge, MA: Harvard University Press, 1998.

Vanderschraaf, Peter. "Covenants and Reputations." *Synthese* 157 (2007): 167–95.

——. "War or Peace? A Dynamical Analysis of Anarchy." *Economics and Philosophy* 22 (2006): 243–79.