The Priority of Social Morality*

Gerald Gaus

"My old mother always used to say, my lord, that facts are like cows. If you stare them in the face hard enough, they generally run away."
~Dorothy L. Sayers, *Clouds of Witness*

1 AN INTRAMURAL DISPUTE (ARISING FROM BOVINE INSPECTION)

Most political philosophers seem to share Mrs. Bunter's view of facts. Facts about motivations, information, as well as social and institutional dynamics, are often seen as pesky cows that need to be stared down so we can get on with spinning out intuitions about true normativity, natural rights or ideal justice, and exchanging those contrived stories (invoking "intuitions") at which philosophers excel. Russell Hardin has long battled this absurd method of political philosophy, which renders so many of its conclusions irrelevant and useless.[1] "The worst failing of contemporary political philosophy is its frequent irrelevance to actual and plausible conditions."[2]

Compared with the median political philosopher, Hardin and I are fellow travellers. We both accept that any adequate view of justice or morality must accommodate the facts of human life, and show how notions of morality and justice facilitate, as well as regulate, myriad forms of human cooperation. And, as Hardin has stressed, questions of scale are critical. In his papers on "Bodo ethics" (more on these anon) he has insisted that systems of moral relations that work well for small-scale closed societies may well be inapplicable to large, impersonal, dynamic, societies.[3] In all this I am a Hardinite, as any reasonable political philosopher should

*Prepared for the Russell Hardin Festschrift, November 6 and 7, New York University.

[1] David Estlund explicitly accepts that the true theory of justice may well have no practical value. See his "Human Nature and the Limits (if Any) of Political Philosophy," *Philosophy & Public Affairs,* vol. 39 (2011): 207-35 and "Utopophobia." *Philosophy & Public Affairs,* vol. 42 (2014): 114-34.

[2] Russell Hardin, "From Bodo Ethics to Distributive Justice," *Ethical Theory and Moral Practice,* vol. 2 (1999): 399–413, at p. 412.

[3] Shaun Nichols and I have argued for this, with special reference to Bodo ethics, in "Moral Learning in the Open Society: The Theory and Practice of Natural Liberty," *Social Philosophy & Policy*, vol. 34 (Spring 2017). See also my essay "Scaling up the Technology of Norm Change: Problems of Justification" at www.gaus.biz.

be.

Just because we share so many assumptions, Russell and I have grounds for fruitful debates. One of these involves the relative roles of instrumental, self-interested, rationality and social morality in explaining human cooperative social life, both small and large. Hardin writes:

> In David Hume's account, our repeated resolution in the same way of an interaction in repetitive contexts may be called a convention that thereafter motivates our coordination with each other. We can commonly see conformance with such conventions as instrumentally rational and self-serving. Some of us might also eventually come to see them as morally binding. *Any such claim of morality must be a later development that comes after the instrumental motivation for following at least some of these conventions*. Even then, the moral motivation may not compel everyone; some might still be compelled primarily by their interest….
>
> The achievement of general social order comes prior to justice, democracy, and other systemic achievements. *It is also prior to any collection of social rules* such as Gaus addresses. These are not about priority in conceptual claims, but in causal claims. Without social order at a relatively high level, we cannot successfully establish and maintain institutions for justice, democracy, and so on.[4]

We need to avoid construing the disagreement between Russell and me as a pointless chicken-and-egg problem. Of course as cooperative orders increasingly secure people's interests, the tendency to comply is increased; but enhanced cooperation raises new problems (including new opportunities for cheating), which then raises new problems of coordination and cooperation that are resolved by the development of social norms and moral rules, which then further enhance the satisfaction of rational interests and allows for further fruitful coordination, and so on and on. In successful cooperative orders there is a virtuous circle between the advancement of the basic interests of participants and normative regulation; it would be folly to suggest that one has absolute causal priority. And, of course, a quintessential convention can evolve into a moral rule.[5] I certainly do not wish to

[4] Russell Hardin, "The Priority of Social Order," *Rationality and Society*, vol. 25 (2013): 407–421, at pp. 407-10, citation deleted, emphasis added. For a different sort of claim that purely instrumental reasoning is in some way more basic than the notion of a rule-based social morality, see Michael Moehler, "The Scope of Instrumental Morality," *Philosophical Studies,* vol. 167 (2014): 431-451.

[5] Moving a bit beyond classic coordination problems, think of a "coordination" interaction such as a Stag Hunt. Even if we have achieved "hunt stag" equilibrium — which might be maintained simply by self-interest — a rule that makes it a moral requirement to hunt stag may stabilize cooperation in the face of trembling hands and other uncertainties.

deny that a convention might pave the way for a rule of what I have called "social morality."

The interesting dispute is, I think, how early in the development of human cooperation guidance by internalized moral rules arises, and what functions it plays when it arises. If internal guidance by what I have called "social-moral rules"[6] is a necessity at the very earliest stages of human cooperation, then, while of course we can accept that other modes of cooperation such as conventions play a role, we should not privilege conventions as somehow "prior" in the development of human social order. The question of how "early" internalized guidance by moral rules arises can take three forms: in how *small of groups*, how *early in the development of moral agents*, and how *early in human history* does internalized guidance by moral rules arise? I believe internalized moral guidance arises very early in all three senses. Normative guidance, shame, cheater detection and punishment are, I believe, fundamental to even the earliest and smallest cooperative orders, and characterizes very young moral agents. Without internalized moral guidance, even small-scale cooperative orders are hopelessly inefficient, and probably impossible.

This paper considers three problems for this view raised by Russell. The next section considers small-scale cooperation; I believe that the evidence indicates that even in very small groups of face-to-face cooperators, the internalization of moral rules is fundamental to their cooperation and cheater suppression. Section 3 then considers Russell's charge that accounts of social cooperation based on moral rules, in which individuals act on the rules despite their interests, are stuck with invoking a variety of somewhat dubious and weak "claims of moral commitment or shared values through [to] Rawls's magical 'addition of the sense of justice and moral

[6] Let us say that for a rule *R* to be a social-moral rule for Betty, Betty must recognize *R* as rule that applies to *C* circumstances; Betty typically has a motivating reason to conform to *R* rather than act simply on her own goals in *C* circumstances on the condition that (*a*) Betty believes that a sufficiently large subset of her group *G* conforms to *R* in *C*; (*b*1) Betty believes that a sufficiently large subset of *G* expects Betty to conform to *R* in circumstances *C* or (*b*2) Betty believes that a sufficiently large subset expects Betty to conform to *R* in *C*, prefers that Betty does so, and will sanction Betty for noncompliance. We also must suppose that *R* is part of a practice of accountability and responsibility, and sustains the moral emotions of guilt, resentment and indignation in relation to violations. See further *The Order of Public Reason* (Cambridge: Cambridge University Press, 2011), pp. 163-181. I am in general following Cristina Bicchieri, *The Grammar of Society* (Cambridge: Cambridge University Press, 2006), p. 11. On accountability and norms, see also Geoffrey Brennan, Lina Eriksson, Robert E. Goodin and Nicholas Southwood *Explaining Norms* (Oxford: Oxford University Press, 2013), especially Part I. In this paper I shall not distinguish the rules of social morality from social norms. They are not, however, equivalent; the rules of social morality are parts of practices of accountability and sustain the moral emotions of guilt, resentment and indignation; not all social norms do so.

sentiment' to make justice work at a large scale."[7] I argue that the evidence in support of internalized rule compliance, even in the face of high costs to personal interests, is impressive, and the underlying mechanisms are not mysterious. Lastly, section 4 briefly turns to the fundamental issue of how social morality functions in large-scale settings and, importantly, whether it is largely displaced by formal legal and political institutions.

2 SOCIAL MORALITY IN SMALL-SCALE SOCIETIES

2.1 Bodo

In several places Hardin has depicted what he has called "Bodo ethics":

> Axel Leijonhufvud… characterizes the village society of eleventh century France in which the villager Bodo lived. We have detailed knowledge of that society from the parish records of the church of St. Germaine. Today one would say that that church is in the center of Paris, but in Bodo's time it was a rural parish distant enough from Paris that many of its inhabitants may never have seen Paris. Virtually everything Bodo consumed was produced by about eighty people, all of whom he knew well. Indeed, most of what he consumed was most likely produced by his own family. If anyone other than these eighty people touched anything he consumed, it was salt, which would have come from the ocean and would have passed through many hands on the way to St. Germaine, or it was spices, which would have traveled enormous distances and passed through even more hands.[8]

In different contexts Russell has focused on different features of Bodo ethics. For present purposes the proposed underlying motivation is of interest:

> A striking feature of Bodo ethics is that it is relatively easily enforceable by the community. *An individual need not rely on self-regulation to be moral*. The knowledge that the whole community has of each individual's adherence to the local moral code allows community members to sanction miscreants. An enormous part of the debate about morality in the modern secular world is about how individuals can be motivated to act morally. That question is answered easily for Bodo's world. *The community spontaneously enforces its morality*

[7] Russell Hardin, *David Hume: Moral and Political Theorist* (Oxford: Oxford University Press, 2007), p. 96.

[8] Russell Hardin, "From Bodo Ethics to Distributive Justice," pp. 401-2. See also Hardin, "The Priority of Social Order," pp. 411ff; Hardin, *Indeterminacy and Society* (Princeton: Princeton University Press, 2003), p. 98.

> *as a set of compulsory norms*. …. The exaction would typically be quick and aimed at the right person.[9]

In this passage Hardin seems to advance what we might call

> *The External Moral Rules Thesis:* In a small cooperative group *G*, a system of social regulation that is seen by members of *G* as simply an *external system of moral rules* is apt to constitute an effective framework for social cooperation.

To be more precise: suppose that *G* is a cooperative group between 20 and 100, in which social regulation is achieved simply through moral rules of type *E* that are generally observed (and publicly known to be) such that (*i*) members of *G* expect the typical member, Alf, to conform to *E*; (*ii*) Alf recognizes that other members of *G* expect him to conform to *E* and will usually punish Alf for infractions of *E*; yet (*iii*) Alf's only motivation for compliance with *E* is self-interest, including the fear of punishment. According to the External Moral Rules Thesis *G* is apt to secure effective cooperation and social order. Note that external moral rules can be, but need not be, rules specifying a classic convention, as punishment may be necessary to secure compliance.[10]

I believe that we have strong evidence that the External Moral Rules Thesis is false. From the very beginning of human social cooperation, social order fundamentally relied on moral rules internalized by the participants. Thus, I shall argue that it is false that "In Bodo's world we do not need morality to keep us all in line because the transparency of all our actions is virtually total."[11]

2.2 Cephu

We now possess rich ethnographic data about rules in small-scale societies. Christopher Boehm has engaged in a massive study of rules and sanctioning practices of both tribal societies and hunter-gather societies.[12] The latter — small groups of 20 to 30 people — is especially interesting for us. Boehm has developed a

[9] Hardin, "The Priority of Social Order," p. 412. Emphasis added.

[10] See, however, the wider characterization of a convention in Samuel Bowles and Herbert Gintis, *A Cooperative Species: Human Reciprocity and its Evolution* (Princeton: Princeton University Press, 2011), p. 111.

[11] Hardin, "The Priority of Social Order," p. 412.

[12] In particular his *Hierarchy in the Forest: The Evolution of Egalitarian Behavior* (Cambridge, MA: Harvard University Press, 1999) and *Moral Origins: The Evolution of Virtue, Altruism and Shame* (New York: Basic Books, 2012).

database of over 300 hunter-gatherer societies and, of these, he has identified about half as essentially closed societies, with minimal contact with agricultural or commercial societies. These societies share much social context of Bodo's village (traditional, small, face-to-face, largely isolated[13]) except, crucially, they are not agricultural and sedentary, and much less hierarchical.

Boehm's data indicates that such small-scale societies tend to employ a hierarchy of punishments, from gossiping and criticism, ridicule, ostracism to capital punishment. Boehm observes that, although "under the spell of Durkheim" anthropologists often depict punishment in small-scale societies as spontaneous and almost automatic, this seems mistaken. Focusing on the sanctioning of overly assertive would-be dominant individuals, Boehm holds that the typical process is considerably more political:

> First, individuals begin to grope toward a group resolution of the problem, initially by gossiping behind the deviant's back and carefully watching the reactions of others. Once consensus seems predictable, some individual still has to lead the sanctioning — unless several group members do so in concert, which can be the case with ridicule. Once in a while the deviant will be simply too intimidating — or too unpredictable — for any one person or even a small coalition to risk taking the first step. [14]

These political dynamics are striking in Colin Turnbull's famous case of Cephu, the cheating hunter. The Pygmy hunters studied by Turnbull sometimes hunt small game with nets. The men place their nets in a long semi-circle, and women and children drive game into the nets. Cephu, having complained of consistent bad luck in hunting, decided to secretly put his nets in front of the others, so game would be first driven into his net. This worked in increasing his take but, unfortunately for him, he was observed. Turnbull continues the account as the hunters

> strode into camp with glowering faces and threw their nets on the ground outside their huts. Then they sat down, with their chins in their hands, staring into space and saying nothing. The women followed, mostly with empty baskets, but they were by no means silent. They swore at each other, they swore at their husbands, and most of all they swore at Cephu….
>
> I tried to find out what had happened, but nobody would say. Kenge, who had been

[13] Whether they are closed in open to debate: much depends on what is meant by this description. Marriage networks, for example, can make the group much more porous than first inspection would indicate.

[14] Boehm, *Hierarchy in the Forest*, p. 118.

sleeping, came out of our hut and joined the shouting. He was the only male who was not sitting down, and although he was young he had a powerful voice, and a colorful use of language. I heard him saying, "Cephu is an impotent old fool. No, he isn't, he is an impotent old animal—we have treated him like a man for long enough, now we should treat him like an animal. Animal!" He shouted the final epithet across at Cephu's camp, although Cephu had not yet returned.

The result of Kenge's tirade was that everyone calmed down and began criticizing Cephu a little less heatedly, but on every possible score: The way he always built his camp separately, the way he had even referred to it as a separate camp, the way he mistreated his relatives, his general deceitfulness, the dirtiness of his camp, and even his own personal habits….

…

Trying not to walk too quickly, yet afraid to dawdle too deliberately, he [Cephu] made an awkward entrance. For as good an actor as Cephu it was surprising. By the time he got to the *kumamolimo* everyone was doing something to occupy himself — staring into the fire or up at the tree tops, roasting plantains, smoking, or whittling away at arrow shafts. Only Ekianga and Manyalibo looked impatient, but they said nothing. Cephu walked into the group, and still nobody spoke. He went up to where a youth was sitting in a chair. Usually he would have been offered a seat without his having to ask, and now he did not dare ask, and the youth continued to sit there in as nonchalant a manner as he could muster. Cephu went to another chair where Amabosu was sitting. He shook it violently when Amabosu ignored him, at which he was told, "Animals lie on the ground."

….

Cephu knew he was defeated and humiliated. Alone, his band of four or five families was too small to make an efficient hunting unit. He apologized profusely, reiterated that he really did not know he had set up his net in front of the others, and said that in any case he would hand over all the meat. This settled the matter, and accompanied by most of the group he returned to his little camp and brusquely ordered his wife to hand over the spoils. She had little chance to refuse, as hands were already reaching into her basket and under the leaves of the roof where she had hidden some liver in anticipation of just such a contingency. Even her cooking pot was emptied. Then each of the other huts was searched and all the meat taken. Cephu's family protested loudly and Cephu tried hard to cry, but this time it was forced and everyone laughed at him. He clutched his stomach and said he would die; die because he was hungry and his brothers had taken away all his food; die because he was not respected.

…From Cephu's camp came the sound of the old man, still trying hard to cry, moaning about his unfortunate situation, making noises that were meant to indicate hunger. From our

> own camp came the jeers of women, ridiculing him and imitating his moans.[15]

In some ways Cephu might seem a good candidate for Russell's model — Boehm muses that he may have been something of an amoral psychopath.[16] But note first that the group decides whether a violation has occurred. Often the lead is taken by one individual, in this case Kenge, who is not necessarily the directly injured party. This helps insure that the dispute will not simply be seen a dyadic conflict. Consensus forms that a violation has occurred; note especially that while Cephu's family does not join in the punishment, neither do they resist. Because small-scale societies are a complex mix of kin and non-kin relations, and it is important that punishment does not lead to inter-family conflict. This is especially clear in cases of capital punishment, which is practiced in many hunter-gather societies.[17] In cases of capital punishment, the entire group of males, including the victim's kin, sometimes collectively kills the offender (in one noted case, the entire group, including women, participated in the execution). In many cases a kin of the offender is selected as executioner.[18] The critical point here is that because eruption of counter-sanctioning is always a possibility, the rule enforced must be seen by all as legitimate, it must be agreed that a violation has occurred, and the kin of the deviant must at least passively accept, and sometimes must actively participate, in the punishment. Lethal weapons abound in hunter-gather groups, and the escalation of violence is an ever-present threat.

As Samuel Bowles and Herbert Gintis more generally stress, effective punishment depends on legitimacy: unless those to be punished and their friends and allies are convinced that the rule being enforced is a legitimate one and one for which community enforcement is appropriate, a punishing action taken as a means to protect social cooperation can lead to weakening it.[19] Experimental evidence

[15] Colin M. Turnbull, *The Forest People* (New York: Simon and Schuster, 1963), pp. 104-8.

[16] Boehm, *Moral Origins*, pp. 44-45.

[17] Boehm reports that in his database about half the hunter-gather societies coded practice capital punishment; there is strong reason to think that the number may be much higher, as central governments treat band and tribal executions as murder. Ibid., p. 84.

[18] Boehm, *Hierarchy in the Forest*, pp. 81-82, 121-22, 180. While females seldom participate in the executions, they do typically participate in the deliberation leading to execution.

[19] Bowles and Gintis, *A Cooperative Species*, p. 26. As Bowles and Gintis point out, in large-scale societies too, anti-social punishment (counter-punishment) is real: experiments show great differences in societies to the extent to which punishment is accepted or evokes counter-sanctioning. (Ibid.) As we shall see in section 3, in experiments in "Power-to-Take" games, Takers who were sanctioned by their partners for taking the partner's endowments but who did not see these takings as unfair, did not decrease their takings in a second round; in contrast, those who

confirms that attempts at punishment readily evoke counter-punishment when the offender does not experience guilt.[20]

2.3 The Internalization of Moral Rules

Note that with Cephu the admission of guilt preceded the group's confiscation of his kill. Consensus on the lower levels of punishment, ridicule and mild ostracism were reached during the walk home and afterwards, and it is this less dangerous level of punishment that triggered his profuse apologies — and only after that did confiscation occur. Still, one might think, all this remains consistent with the External Moral Rules Thesis. After all, it was punishment that in the end drove Cephu to admit guilt, and Cephu was known to be something of an actor, so his profuse admissions of guilt may simply have been strategic.

The important point, though, is how costly such punishing episodes are to the group. Hunting is a highly egalitarian, cooperative, activity and shirkers, cheats, and free-riders such as Cephu pose real threats. Cephu, indeed, not only posed the threat of a cheat, but he initially resisted punishment and sought to intimidate others, arguing that he was an important person, indeed a chief.[21] Cephu, perhaps, did view the rules largely externally, and that is why he was a persistent problem. Rules that were generally perceived as purely external by group members, depending solely on self-interest to motivate compliance, would be a hopelessly inefficient way of securing cooperation, inviting both opportunistic evasion and counter-punishment. The large majority must, and do, internalize the rules, which, as Boehm rightly says, involves emotional attachment to the rules and compliance with them.[22] Such individuals have a virtue highly prized in many small hunter groups — self-control.[23] In the face of temptations to cheat and dominate, they can be counted on to generally comply with the group's rules. Cephu was lacking in self-control and was a severe problem for the group: he needed watching. Those even more seriously lacking in self-control, such as repeated murderers, can be executed.[24] Overall,

---

were sanctioned and did think their initial taking unfair (but hoped to get away with it) responded to sanctioning by decreasing their takings.

[20] Astrid Hopfensitz and Ernesto Reuben, "The Importance of Emotions for the Effectiveness of Social Punishment," *The Economic Journal*, vol. 119 (October 2009): 1534–1559.

[21] Boehm, *Moral Origins*, p. 43

[22] Ibid., pp. 113-14.

[23] For a striking case, see Boehm, *Hierarchy in the Forest*, pp. 51-59.

[24] In Boehm's database, of the societies that engaged in capital punishment, repeat murder was the second most reported capital offense.

Boehm argues, hunter-gather societies display a high level of rule internalization and corresponding self-control.

Students of cognition have recently turned to modeling the processes that underlie norm internalization.[25] We know that internalization of moral rules is a normal accomplishment for humans, and occurs at a very young age. In a series of experiments conducted by Gertrude Nunnar-Winkler and Beate Sodian, children between four and eight were told a story about two children, both of whom liked candy. The first child was tempted to steal the candy, but did not; the second stole the candy. Even the four-year-old subjects knew that stealing was wrong and could provide reasons why this is so. Thus they could engage in punishing violators. The difference is that the youngest children expected the child who stole the candy to be happy with his violation of the rule, while they (the youngest children) expected the child who resisted temptation to be sad. Older children reversed this; they supposed the child who stole would be sad – guilty – while the child who resisted temptation would be the happy one. Younger children apparently expect people to be happy when they get what, all things considered, they want, regardless of whether this violates a moral requirement and harms others.[26] Again, older children expected the violator to feel unhappy. Nunnar-Winkler and Sodian conclude:

> children may first come to know moral rules in a purely *informational* sense, that is, they know that norms exist and why they should exist. Not until several years later, however, do they seem to treat them as personally binding obligations the intentional violation of which will be followed by negatively-charged self-evaluative emotions or genuinely empathetic concerns.[27]

Very young children view moral rules as external guides, as in the External Moral Rules Thesis. They can appreciate reasons that these rules are important and even that punishment is appropriate; what they do not grasp is that the rule can function as a requirement in an agent's deliberations and can be seen as "personally

[25] See Giulia Andrighetto, Daniel Villatoro and Rosaria Conte, "Norm Internalization in Artificial Societies," *AI Communications*, vol. 23 (2010): 325–339.

[26] It is generally thought that young children see harm to others as violating a basic moral requirement. See Elliot Turiel, Melainie Killen, and Charles C. Helwig, "Morality: Its Structure, Functions and Vagaries," in *The Emergence of Morality in Young Children*, edited by Jerome Kagan and Sharon Lamb (Chicago: Chicago University Press, 1987): 155–243 at p. 174. Guilt is especially associated with violation of rules against harm and the rights of others. Jesse Prinz, *The Emotional Construction of Morals* (Oxford: Oxford University Press, 2007), p. 77.

[27]Nunner-Winkler and Sodian, "Children's Understanding of Moral Emotions," *Child Development*, vol. 59 (October 1988): 1323–38 at p. 1336. Emphasis in original.

binding,"[28] so that the agent will feel guilt for failing to meet this requirement even if by so doing she gets what she wants. What very young children do not grasp is that a typical moral agent cares about moral requirements and so can put aside the things that she wants and, instead, conform to the rule's requirements, and success in doing this relates to her own self-esteem. As Abraham Lincoln was said to have remarked, "when I do good, I feel good. When I do bad, I feel bad. That is my religion."[29]

2.4 What's So Special about Hunter-Gather Societies?

I have focused on contemporary hunter-gatherer societies (with some reference to larger tribal societies), whereas Hardin's "Bodo" resided in a medieval agricultural community. In trying to think about Russell's question of the "priority" of social order *v*. social morality, which is the better model? I believe the generally accepted answer is that humans evolved our technology of social cooperation within such hunter-gatherer bands, and so if our concern is some sense of priority, then it is these bands that formed the context of the evolution of human cooperation.

Just when, and why, our human ancestors became intense cooperators, is of course disputed, and so any claims we make must be highly tentative (that's the feature of facts that leads so many philosophers to try to stare them down). It is clear that humans have long been engaged in deeply cooperative hunting. Mary Stiner and her colleagues discovered distinctive differences in the bones of the carcasses of human kills between 400,000 and 200,000 years ago at Qesem Cave in Israel. Bones from carcasses from 400,000 years ago demonstrate that the human hunters employed tools to cut the meat, but the cut marks indicate the presence of a number of different cutting implements employed at different angles. Evidence from this earlier period suggests that

> meat distribution systems were less staged or canalized than those typical of Middle Paleolithic, Upper Paleolithic, and later humans. The evidence for procedural interruptions and diverse positions while cutting flesh at Qesem Cave may reflect, for example, more hands (including less experienced hands) removing meat from any given limb bone, rather

[28] Ibid., p. 1324.

[29] See Bowles and Gintis, *The Cooperative Species,* p. 169. Bowles and Gintis devote much care to analyzing how internalization of social morality can be modeled (chap. 10). As they stress, the internalization of norms is an aspect of cultural transition that affects preferences or values. On the general phenomenon of cultural transmission, see Peter J. Richerson and Robert Boyd, *Not by Genes Alone: How Culture Transformed Human Evolution* (Chicago: University of Chicago Press, 2005).

> than receiving shares through the butchering work of one skilled person. Several individuals may have cut pieces of meat from a bone for themselves, or the same individual may have returned to the food item many times. Either way, the feeding pattern from shared resources may have been highly individualized, with little or no formal apportioning of meat.[30]

Kills from 200,000 years ago display much more uniform cut marks, indicating a single cutter, who cut and distributed the kill. A very plausible hypothesis that by this time humans were, or were well on their way to becoming, distinctly egalitarian hunters. Distribution of the kill does not seem, as in the earlier case, determined by competition among the hunters (where we can suppose the more dominant took the best, first), but by a designated cutter allocating shares of the kill (as is the case in many contemporary hunter-gather societies). To be a bit more speculative, it looks as if the socialized primate carnivores of 400,000 years ago were becoming egalitarian hunters by 200,000 years ago. It is very difficult not to conclude that egalitarian sharing of cooperative hunts had already taken root by this period. Self-control was absolutely essential to the development of such egalitarian sharing.

We have good reason to conclude that modern, late-Pleistocene, humans lived in groups of between 25 and 150,[31] obtained a high percentage of their calories from hunting or fishing, and engaged in egalitarian meat sharing. Boehm's central thesis is that the mode of life of our common cooperative ancestors is essentially that of today's hunter-gather societies. As I have remarked, in his important study of contemporary late-Pleistocene-appropriate ("LPA") foraging societies, Boehm eliminated from consideration societies that have been heavily influenced by Western and market societies, those with some agriculture, those that trade with agricultural groups, those that rely on domesticated horses, and so on, ultimately identifying 150 (of which a third have been more minutely analyzed) contemporary forger societies whose way of life corresponds to what we know of late-Pleistocene hunter-gatherer bands.[32]

[30] Mary C. Stiner, Ran Barkai, Avi Gopher and James F. O'Connell, "Cooperative Hunting and Meat Sharing 400–200 KYA at Qesem Cave, Israel," *Proceedings of the National Academy of Sciences of the United States of America*, 106, No. 32 (Aug. 11, 2009): 13207-13212 at 13211.

[31] Daniel Friedman points to 150, with much larger numbers when groups fused. *Morals and Markets: An Evolutionary Account of the Modern World* (New York: Routledge, 2008), p. 16. See also David C. Rose, who mentions 200 as the typical size of the groups in which humans evolved; *The Moral Foundations of Economic Behavior* (New York: Oxford University Press, 2011), chap. 3. Closer examination shows that group size may be understood differently: average band size may differ from typical group size. See Bowles and Gintis, *The Cooperative Species*, p. 95.

[32] Boehm, *Moral Origins,* pp. 78-82.

This assumption is certainly not uncontroversial.[33] Contemporary LPA-foraging societies exist in the Holocene era of much, much, milder climates and arguably greater ease, or at least less uncertainty, in obtaining food. In the extraordinarily harsh late-Pleistocene climate, it could well have been far less rare for groups to have faced such dire circumstances that sharing broke down, leading to the group splintering into family-sized, rather than band-sized, units, with very different evolutionary dynamics.[34] Nevertheless, the social organization of these societies corresponds to much of what we know about late-Pleistocene bands — they are mobile, stress sharing rather than storing meat, combine hunting with foraging and live in core bands of 20 to 30 persons. And some of these current LPA societies have, like late-Pleistocene bands, faced the most dire of circumstances — leading in some cases to parents eating their children.[35] At present, I believe, our best estimates of the earliest form of intense human cooperative social orders correspond to these "LPA-appropriate" hunter-gather societies, and these societies are ones in which, while punishment is a critical form of social control, it must be used carefully, its dangers mitigated by the internalization by most members of the group's rules and their self-control in the form of conscience. To put the matter bluntly: given our best current information, the evolution of social order marched hand-in-hand with the evolution of internalized social morality or, as Kitcher puts it, "normative guidance."[36]

To be sure, given the incredibly swift cultural evolution of the last 10,000 years[37] we cannot assume that our current social morality is anything similar to the egalitarianism of LPA societies.[38] The point however, is that the normative competencies we find in such societies — such as norm internalization and its attendant motivation — are almost surely long-standing features of human social cooperation. So far from being odd commitments of confused, obscurantist, Kantian

[33] For doubts, see Peter J. Richerson and Robert Boyd, "Rethinking Paleoanthropology: A World Queerer than We Supposed," in *Evolution of Mind,* edited by Gary Hatfield and Holly Pittman (Pennsylvania Museum Conference Series, 2013): 263-302.

[34] Boehm, *Moral Origins,* pp. 274ff. On the other hand, it could well have been such instability that increased the benefits of cooperation. See Bowles and Gintis, *The Cooperative Species*, pp. 93ff.

[35] Boehm, *Moral Origins,* p. 275.

[36] Philip Kitcher, *The Ethical Project* (Cambridge, MA: Harvard University Press, 2010), chap. 2.

[37] To what extent genes have evolved during this period is a highly controversial question. 10,000 years is far less than the 1000 generations, which is the rule-of-thumb for the evolution of major traits. But this is a highly controversial matter that is being debated.

[38] Though I have argued that it is surprisingly so. "The Egalitarian Species," *Social Philosophy and Policy*, vol. 31 (Spring 2015): 1-27.

philosophers,[39] they are universal features of cooperating groups of humans.

3 NORMATIVE COMMITMENT AND SENSITIVITY TO RULES

3.1 How Muscular is Normative Commitment?

There is, then, nothing truly mysterious about a deeply cooperative species internalizing, and so becoming emotionally attached to, the rules that specify the terms of social cooperation, such as moral rules concerning sharing and property. This was probably fully accomplished 45,000 years ago in small groups. Having been hard on the modal political philosopher, in fairness I must observe the characteristic blind spot of many PPE-oriented philosophers, who accord an almost religious status to the manifestly false axiom that rationality concerns something like a pursuit of self-interested goals.[40] This is entailed by neither the idea of instrumental rationality nor rational choice/decision theory, and even a cursory understanding of moral psychology displays its deep implausibility. But like many widely-accepted false claims there is a genuine insight lurking here — the entirely sensible worry that such moral rule based motivations may not be able to stand up to significant temptations to pursue one's narrow interests by defecting. "A mere norm," Hardin writes, "is unlikely to override self-interest in many such contexts. Some members might be sufficiently motivated by moral commitments, but we cannot generally expect everyone to be, especially when the stakes are high."[41] So, accepting that normal humans internalize, care about and are motivated to conform to, social morality, one may well wonder whether such merely "normative" motivation can successfully hold up to self-interest.

3.2 Social-Moral Rule Sensitivity

Cristina Bicchieri has usefully modeled this problem in terms of norm sensitivity.[42] Sensitivity to a norm or social rule concerns the relation between the content/function of the rule and the moral and value commitments of a person. When a rule of social morality is strongly supported by an agent's own normative

[39] Ken Binmore, *Natural Justice* (Oxford: Oxford University Press, 2005), pp. vii-viii.

[40] Thus the common depiction of Hobbes as somehow the father of rational choice theory (even though Hobbes himself had a much more sophisticated view of human motivation). See for example, Hartmut Kliemt, *Philosophy and Economics I: Methods and Models* (Munich: Oldenbourg, 2009), pp. 46ff.

[41] Hardin, "The Priority of Social Order," p. 414.

[42] Cristina Bicchieri, *The Grammar of Society,* p. 62.

commitments, she will tend to be highly sensitive to a norm: put simply, she has many reasons for adhering to the requirements of a norm even in the face of temptations to cheat based on narrow self-interest.[43] As one's personal normative commitments and beliefs provide less support for the norm, sensitivity will decrease. A person whose only reason for compliance is fear of punishment would, on this view, tend to have a low sensitivity: he will engage in opportunistic cheating behavior when he can get away with it, or when the expectations of gain outweigh the likely punishment. Thus we can hypothesize:

> *The Justification Effect:* Alf's sensitivity to a rule of social morality tends to rise as its justification to Alf increases, where justification depends on the coherence of the rule with Alf's personal normative beliefs and convictions.

Bicchieri is clear that (what I have called) the Justification Effect varies in the population. Those with greater "reflective autonomy," she predicts, will have a stronger tendency to decrease their sensitivity to a norm as they become aware of reasons against it, while more conformist members of the group will have higher sensitivity to a rule just because, say, it has been in place for a long time, and will be less sensitive to reasons against it.[44] On the other hand, as I have said, those whose sole reason to act on the norm is the fear of punishment will have much less sensitivity to the norm and will be open to opportunistic cheating.[45]

The Justification Effect shows the importance of what I have elsewhere called "convergent normativity."[46] Many political philosophers are apt to think of the entire notion of "public reason" as a mere piece of Rawlsian jargon — and, alas, too often it is. However, it also allows us to see a fundamental feature of an effective social morality. As the rules of social morality tend toward public justification in group *G*, in the sense that overwhelmingly the members of *G* find that their personal normative beliefs and convictions support the rules, the members of *G* became more sensitive to those rules. And, so, internal motivations for compliance are stronger,

[43] On the importance of reasons, see and Bicchieri and Hugo Mercier. "Norms and Beliefs: How Change Occurs" in *The Complexity of Social Norms*, edited by Maria Xenitidou and Bruce Edmonds (New York: Springer, 2014): 37-54.

[44] See her forthcoming *Norms in the Wild*.

[45] Plausible models of internalization often yield polymorphic results, with a population divided between internalizers and more opportunistic types. See for example Andrighetto, Villatoro and Conte, "Norm Internalization in Artificial Societies."

[46] See, for example, my *Tyranny of the Ideal: Justice in a Diverse Society* (Princeton: Princeton University Press, 2016), chap. IV.

and socially costly and perhaps disruptive acts of punishment can be reduced. The more individuals find that the rules they live by correspond to their important personal values, moral and religious convictions, the more they are inclined to follow these rules even in cases when the rules call for significant sacrifice of their narrow interests.

This is not to deny the basic truth that, as Peter Richerson and Robert Boyd put it, "we are imperfect and often reluctant, though often very effective cooperators."[47] We need moral rules because we are a complex combination of selfish and cooperative creatures: the moral system, we might say, has developed on top of an earlier selfish set of motivations.[48] Nevertheless, this moral system is real, and is a critical basis of the human cooperation. When it draws on the personal values and moral convictions of the participants, their motivational power can be channeled into social morality.

3.3 The Puzzle of Punishment

Nevertheless punishment is necessary for an effective system of social morality. Some, such as Cephu, may only be sensitive to the rules insofar as they expect punishment. More common is to have modest sensitivity, willing to abide by the rules but not at great costs, while many others have quite high sensitivity. But even they are usually concerned with self-interest, and seek ways to advance it. Boehm hypothesizes that we evolved a "flexible conscience" — able to distinguish what truly must not be done from minor violations and "exceptions" that allow us wiggle room to advance self-interest.[49] But appealing to punishment is no quick solution to our problem. Why do people bother to punish? To be sure, in iterated interactions based on direct reciprocity, "punishing" acts are actually elements of an optimizing strategy, and so enhance the interests of the punisher.[50] While direct reciprocity can be effective in accounting for cooperation in very small groups (dyads, triads) its

[47] Peter J. Richerson and Robert Boyd, "The Evolution of Free Enterprise Values," in *Moral Markets: The Critical Role of Values in the Economy*, edited by Paul Zak (Princeton: Princeton University Press, 2008), p. 114.

[48] Ibid. See also Freidman, *Morals and Markets*, chap. 1.

[49] Boehm, *Moral Origins*, pp. 172-78. Bicchieri and her co-workers have shown how subjects exploit normative ambiguity in order to provide wiggle room to advance their interests. See Bicchieri and Alex Chavez, "Norm Manipulation, Norm Evasion: Experimental Evidence," *Economics and Philosophy,* volume 29 (Special Issue 2, July 2013): 175-98. Bicchieri and Hugo Mercier, "Self-serving Biases and Public Justifications in Trust Games," *Synthese,* vol. 190 (2013): 909–922.

[50] As in the folk theorem, stressed by Ken Binmore, *Natural Justice*, chap. 5.

capacity to sustain cooperation dramatically decreases as group size increases.[51] Bodo's group of 80 would be far too large for direct reciprocity to sustain cooperation.[52] In large groups one who punishes an infraction is also following a moral rule at the cost of her own interests — she would almost surely be better off ignoring the infraction and go about her own business. So while it is certainly true that punishment is necessary to sustain a cooperative morality, it simply pushes us back to the question: why do people support morality by punishing rather than allow the infraction to pass?

We clearly do have a keen capacity to detect cheaters. And, like internalization, cheater detection is a very early human accomplishment, being manifest in 3- and 4-year olds.[53] And it is a capacity we effectively employ: as extensive empirical research has demonstrated, people do punish, and often at significant-to-high costs to themselves.[54] To focus on a very familiar case, in Ultimatum Games Responders will often refuse sizable stakes, and walk away with nothing rather than accept miserly offers.[55] Bicchieri has effectively argued that underlying this behavior is a

[51] See Natalie Henrich and Joseph Henrich, *Why Humans Cooperate* (Oxford: Oxford University Press, 2007), p. 51. Indirect reciprocity, or reputation, might seem to underwrite cooperation in larger groups by encouraging "boycotts" of violators, but indirect reciprocity turns out to be very sensitive to the quality of information about people. See Henrich and Henrich, *Why Humans Cooperate*, chap. 4; Bowles and Gintis, *The Cooperative Species,* pp. 68-70; Peter Vanderschraaf, "Covenants and Reputations," *Synthese,* vol. 157 (2007): 167–95.

[52] In Bowles and Gintis's agent-based modeling allowing even for small rates of errors in reciprocation, groups over 10 seldom, and over 15 essentially never, evolved cooperation. *The Cooperative Species*, pp. 64-68. Even in small group forager bands, direct reciprocity does not explain most cooperation. Boehm, *Moral Origins*, pp. 179-80.

[53] See Denise Dellarosa Cummins, "Evidence for the Innateness of Deontic Reasoning," *Mind & Language*, vol. 11 (June 1996): 160–90; "Evidence of Deontic Reasoning in 3- and 4-year-olds." *Memory and Cognition*, vol. 24 (1996): 823–829.

[54] The experimental work on strong reciprocity and altruistic punishment is now extensive. The pioneering work was done by Ernst Fehr and his colleagues. See, for example, Fehr and Urs Fischbacher. "The Economics of Strong Reciprocity" in *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, edited by Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr (Cambridge, MA: MIT Press, 2005): 151–91; Fehr and Simon Gächter, "Cooperation and Punishment in Public Goods Experiments, *American Economic Review*, vol. 90 (Sept. 2000): 980–94; Fehr and Gächter, "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, vol. 14 (Summer 2000): 159–81.

[55] The now famous Ultimatum Game is a single-play game between two anonymous subjects, Proposer and Responder, who have *X* amount of some endowment (say, money) to distribute between them. In the classic version of the game, Proposer makes the first move, and gives an offer of the form, "I will take *n* percent of *X*, leaving you with 100–*n* percent," where *n* is not greater than 100 percent. If Responder accepts, each gets what Proposer offers; if Responder rejects, each receives nothing. For a recent overview see Eric van Damme et. al., "How Werner Güth's Ultimatum Game Shaped Our Understanding of Social Behavior," *Journal of Economic Behavior & Organization*, vol. 108 (2014): 292–318.

concern with fairness norms.[56] To recall the familiar: in the United States and many other countries, one-shot Ultimatum Games result in median offers of Proposers to Responders of between 50 percent and 40 percent, with mean offers being 30 percent to 40 percent. Responders refuse offers of less than 20 percent about half the time.[57] Play in Ultimatum Games does not importantly differ by gender or age. And, importantly for our purposes, Responder rejection rates remain high even when stakes are significantly increased. A variety of studies have shown that play in Ultimatum Games is not highly sensitive to the absolute size of the endowments being divided. In some studies raising the stakes from, say $10 to $100 typically have no significant effect.[58] These are common results.[59] However, although Responder rejection rates remains high even when playing for surprisingly high amounts, raising the stakes eventually does have the effect of decreasing rejection rates (Responders end up taking low offers rather than going away with nothing). As Steffen Andersen and his co-researchers point out, in many Ultimatum Game experiments Proposers advance very few low offers, making it difficult to judge what Responders would do in the face of such offers. In their recent study, some treatments drastically increased the size of endowments to be divided (equivalent to 1,600 hours of work in India, where the experiment took place) and they elicited many low offers by Proposers. In treatments with traditional sized stakes the behavior of Responders was in line with normal play (though there were more low offers to be rejected); in their very high stakes treatments only 1 of 24 Responders rejected low offers.[60]

[56] Bicchieri, *The Grammar of Society*, chap. 3.

[57] Ibid., p. 105. Here some small-scale societies are outliers. See Joseph Henrich and Natalie Smith. "Comparative Evidence from Machiguenga, Mapuche, and American Populations" in J. Henrich, R. Boyd, S. Bowles, et al. eds., *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societie*s (Oxford: Oxford University Press, 2004): 125–67.

[58] See Elizabeth Hoffman, Kevin A. McCabe and Vernon L. Smith, "On Expectations and the Monetary Stakes in Ultimatum Games," *International Journal of Game Theory,* vol. 25 (1996): 289-301.

[59] See, e.g., Robert Slonim and Alvin E. Roth, "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic," *Econometrica*, vol. 66, No. 3 (May, 1998), pp. 569-596. In one study with an endowment worth three months wages still displayed Responder rejection of lower offers. Bicchieri, *The Grammar of Society*, p. 114n.

[60] Steffen Andersen, Seda Ertaç, Uri Gneezy, Moshe Hoffman and John A. List, "Stakes Matter in Ultimatum Games," *The American Economic Review*, vol. 101 (December 2011): 3427-3439. See also Slonim and Roth, "Learning in High Stakes Ultimatum Games."

3.4 Reactive Emotions

Stakes do matter in Ultimatum Games, but it typically takes high stakes before low offers are common and commonly accepted. Although we must accept the thesis that motivations based on the internalization of social morality has its limits, in many ways such motivation is surprisingly strong in those who have been on the short-end of the unfairness stick. Why are so many Responders in Ultimatum Games so ready to deprive themselves of significant resources when there is no possibility of compensating gains through future interactions?[61] A hypothesis with strong experimental support is that reactive emotions such as anger are critical in motivating punishing behavior.[62]

Overall, I think we have good reason to accept what I shall call the *Reactive Emotion View*: Responders' rejection of low offers is partly explained in terms of Responders' emotional reaction to the offers Proposers make to *them*,[63] in particular whether the offer evokes negative emotions such as anger, irritation, or envy.[64] General theories of emotion support the anger/irritation/indignation version of this view; as Nico H. Frijda notes, anger and indignation are generally evoked by norm violation.[65] The main idea here is that, in addition to one's sensitivity to the norm,

[61] One possible explanation — one that Russell sees as partaking of the magical — is that people may be moved by a sense of justice [see John Rawls, *A Theory of Justice*, rev. edn (Cambridge, MA: Harvard University Press, 1999), chap. VIII.]. I do not think it is magical, and some evidence indicates that impartial concern for justice may be a motivational factor. Kevin M. Carlsmith, John M. Darley, and Paul H. Robinson, "Why Do We Punish?: Deterrence And Just Deserts As Motives For Punishment," *Journal of Personality and Social Psychology,* vol. 83 (2002): 284-99. Third-party punishment might be seen as based on an impartial sense of justice, and there is certainly considerable evidence for such punishment. See also Ernst Fehr and Urs Fischbacher, "Third-party Punishment and Social Norms," *Evolution and Human Behavior*, vol. 25 (2004): 63 – 87. However, I do not think the evidence indicates this to be a critical factor, once we have factored out the reactive moral emotions, such as anger. In an interesting experiment Simon Knight sought to determine whether Responders were upholding such a sense of justice — whether "the concern is with unfair offers in general" — or were responding not to the Proposer's general status as a sharer or miser, but specifically what the Proposer did to *her* — whether the Proposer gave *her* a high or low offer. Knight finds that Responders' behavior supports the latter hypothesis — that Responder Betty's action is more strongly influenced by what has been done to *her*, so she will be apt to accept a high offer from a generally unfair Proposer or reject one from a generally fair one. See Simon Knight, "Fairness or Anger in Ultimatum Game Rejections?" *Journal of European Psychology Students,* vol. 3 (2012): 1-14.

[62] See, e.g., Hopfensitz and Reuben, "The Importance of Emotions for the Effectiveness of Social Punishment."

[63] Thus my focus at present is second-party, not third-party, punishment.

[64] Ronald Bosman, Joep Sonnemans, Marcel Zeelenberg, "Emotions, Rejections, and Cooling Off in the Ultimatum Game" (2001) at http://hdl.handle.net/11245/1.418488; Kirchsteiger, "The Role of Envy in Ultimatum Games."

[65] Nico H. Frijda, *The Emotions* (Cambridge: Cambridge University Press, 1996), p. 311.

those who are on the receiving end of defection — or, by extension, those who empathize with victims —tend to get angry or irritated, and this makes them less sensitive to the costs of their punishing activities.[66]

According to the Reactive Emotions View, low offers, defined as where *X*–*n* is (1) a small amount and (2) *n* is a large proportion of *X*, should tend to be rejected: the personal costs of rejection are low (*X*–*n* is small) but we would expect an emotional reaction (because *n* is a very large percentage of *X*). Conversely, high offers, where *X*–*n* is a sizable amount and *n* is a small proportion of *X*, should be accepted: the costs of rejection are high and the negative emotional reactions should be low or non-existent (indeed, the Responder may have extra incentive to accept if her reactive emotion is joy at getting so much!). This is the generally observed behavior.[67] But what of offers that are absolutely large, but proportionally low (i.e., in *X*–*n*, *n* is a very high percentage of *X*, but the absolute size of *X*–*n* is large)? As we have seen, although Responder reactions are not highly sensitive to stakes, they do matter: rejection rates go down for very high stakes. This is consistent with the Reactive Emotions View, which depicts a trade-off rate between the costs of punishment and the negative emotions attached to being treated badly.[68] The crux of the Reactive Emotions View is that negative emotions can provide extra incentive to engage in costly punishment, not that the emotional reactions are so strong that even very large gains (say a 20% share of over 1500 hours wages) will be angrily rejected. After all, we would expect that the value of monetary gains will always be increasing, but one can get only so angry: if so, at some point the value from monetary gains curve will intersect negative value of emotional reaction, leading to Responders to accept the offer.

If this is correct, and we suppose that emotions are more subject to fluctuation

[66] Another cost to which punishers appear insensitive is the number of violators; even if defection is "the norm" — there are many defectors — punishment does not generally decrease. Jonathan Bone, Antonio S. Silva and Nichola J. Raihani, "Defectors, Not Norm Violators, are Punished by Third-Parties," *Biology Letters*, 10: 20140388. http://dx.doi.org/10.1098/rsbl.2014.0388.

[67] See, e.g., Knight, "Fairness or Anger in Ultimatum Game Rejections?" pp. 7-8. As we shall see in the next section, expectations count.

[68] To drastically oversimplify, The Reactive Emotion View can be modeled as claiming the decisions are based on a two-part value function. Letting *X*–*n* be an offer in an Ultimatum Game, where *X* is the total endowment and *n* is the percentage that the Proposer reserves for himself, then Responder's total value of the *X*–*n* offer will be $V_{MG} - V_{RE}$, where V_{MG} is the value of the absolute monetary gain, and V_{RE} is the value based on the reactive emotions, a value arising from the negative emotions, which focus on the ratio of *X* to *n*, as mediated by expectations of what is to be expected. A Responder will accept if total value is positive, reject if it is negative.

than the costs of punishing activity (such as forgone monetary gains in the Ultimatum Game) a reasonable hypothesis is that Responders will "cool down" after a time delay. That is, we would expect Responders to accept an offer after a cool down period that they would immediately reject. The results of experiments are mixed, but I believe generally support this hypothesis. In an earlier study a break of an hour had no effect,[69] while the more recent study of Veronika Grimm and Friederike Mengel found a marked decrease in rejection rates after only ten minutes: "While almost no low offers are accepted without delay, a large share (65–75%) of these offers gets accepted after a 10 minutes delay only."[70] Grimm and Mengel also found that low offers of Proposers increase after a break; this is consistent with work on Dictator Games,[71] which indicates that Dictators whose decisions are driven by immediate affect rather than calculation make more generous offers; apparently a cool down period gives each party time to switch into calculation mode, which favors focusing on the forgone personal benefits or incurred personal costs.[72] In an experiment on the related "Power-to-Take Game" (see section 3.5) a more complicated pattern emerged: here both a "cooling off" and a "getting steamed up" effect seemed present. If the Proposer's actions are not too selfish from the perspective of the Responder, the Responder seems to cool off after a wait time; however as Proposers get greedier, wait time *raises* the Responders' level of punishment.[73] If both cooling off and getting steamed up occur, we would expect ambiguous results from wait time experiments.

3.5 Emotions in Power-to-Take Games

A problem with measuring the role of emotions in straightforward Ultimatum Games is that Responders only have a take-it-or-leave-it choice and, as we have seen, low offers are typically uncommon. The role of emotions in Responders' behavior has been extensively studied in a "cousin" of the Ultimatum Game, the Power-to-

[69] Ronald Bosman, Joep Sonnemans, Marcel Zeelenberg, "Emotions, Rejections, and Cooling Off in the Ultimatum Game."

[70] Veronika Grimm and Friederike Mengel, "Let Me Sleep on It: Delay Reduces Rejection Rates in Ultimatum Games," *Economics Letters*, vol. 111 (2011): 113-115.

[71] In the so-called "Dictator Game" Proposer simply decides on the two shares, and that's the end of the game (not much of a game, to tell the truth).

[72] Jonathan F. Schulz, Urs Fischbacher, Christian Thön and Verena Utikal, "Affect and Fairness: Dictator Games under Cognitive Load," *Journal of Economic Psychology*, vol. 41 (2014): 77–87.

[73] Fabio Galeotti, "An Experiment on Waiting Time and Punishing Behavior," *Economics Bulletin*, vol. 33/2 (2013): 1383-1389.

Take Game, which allows more scope for variable emotional reaction. A Power-to-Take Game involves two players, a Taker and a Responder; their roles are determined at random. To start, each player is given an endowment; in some treatments the players earn their endowment in a pre-game task, in others it is simply distributed by the experimenter. Suppose the endowment for each is Y_{Take} and Y_{Resp}. The Taker, then determines take rate — the proportion of the Responder's endowment he will take. The Responder then has an option of destroying any amount of her endowment that she wishes, before the Taker's percentage is transferred from her. So if the endowment was $10, and the Taker's announced a take rate of 50%, the Taker would get $5 if the Responder destroyed none of her endowment, which would yield total payoffs of $15 for Taker and $5 for Responder. If the Responder decides to destroy half her endowment after the Taker announces his take rate, it would reduce her endowment to $5, of which the Taker would get $2.50. This game is sometimes described as an Ultimatum Game that allows variable punishment, since Responder can decide on the level at which she will deny Taker's resources.[74] But note that in this game the Responder cannot affect the Taker's endowment, but only the amount of her endowment the Taker can transfer.[75]

In an early pioneering study by Ronald Bosman and Frans van Winden, where players earned their endowments, out of 39 subjects, only three Takers took 0, positive takings ranged from 25-100%, with a mean of 58.5%, and median 66.7%; 70% was the mode.[76] Eight Responders chose to destroy part of their endowment, and of these, seven destroyed the entire endowment. In a later study Bosman, Matthias Sutter and van Winden compared this play to another experiment in which endowments were simply distributed at the start of play. [77] Play in the no effort experiment was markedly different; Takers took an average of 32% more, and many more Responders destroyed, and more opted for intermediate destruction rates. Display 1 summarizes the differences between the effort and no effort experiments.

[74] The variability of destruction is meant to uncover the relation of degree of emotional response to degree of punishment; I discuss presently a version of Power-to-Take that gives only limited punishment options which, not too surprisingly, considerably blunts the importance of emotions.

[75] See Ernesto Reuben and Frans van Winden, "Fairness Perceptions and Prosocial Emotions in the Power To Take," *Journal of Economic Psychology*, vol. 31 (2010) 908–922 at 908.

[76] Ronald Bosman and Frans van Winden, "Emotional Hazard in a Power-to-Take Experiment," *The Economic Journal*, vol. 112 (January 2002): 147-169 at p. 153. This is typical of takings in Power-to-Take Games; see Reuben and van Winden, "Fairness Perceptions and Prosocial Emotions in the Power To Take," p. 912.

[77] Ronald Bosman, Matthias Sutter and Frans van Winden, "The Impact of Real Effort and Emotions in the Power-To-Take Game, " *Journal of Economic Psychology*, vol. 26 (2005): 407–429.

	Effort	*No Effort*
Destroy Everything	7	6
Destroy Part	1	9
Destroy Nothing	31	25
Total	39	40

DISPLAY 1: RESULTS IN TWO POWER-TO-TAKE EXPERIMENTS[78]

Especially interesting is that these experiments sought to determine the extent to which emotional reactions explained behavior. Emotions were measured via self-reporting on a seven-point scale ranging from ''no emotion at all'' (1) to ''high intensity of the emotion'' (7). The emotions measured were irritation, anger, contempt, envy, jealousy, sadness, joy, happiness, shame, fear, and surprise.[79] The following findings are of interest to us:

- Responders who destroyed report more intense emotional reactions than those who do not.
- The most intense emotions of Responders who destroy in the *no effort* condition were (in order) anger, contempt, surprise and irritation.
- The most intense emotions of Responders who destroy in the *effort* condition were (in order) irritation, contempt, surprise and anger; the emotions tended to be more intense in this treatment.

[78] Ibid., 418.

[79] Ibid., 415. "In both conditions, the sequence of actions was as follows. Before subjects played the one-shot PTT-game, they were randomly divided into two groups. One group was referred to as participants A (the take authorities) and the other as participants B (the responders). Subsequently, random pairs of a responder and a take authority were formed by letting take authorities draw a coded envelope from a box. The envelope contained a form on which the endowment of both participant A and participant B was stated. The take authorities then had to fill in a take rate and put the form back in the envelope again. After the envelopes were collected, we asked the take authorities to report their emotions as well as their expectation of what the responder would do. The envelopes were brought to the matched responders who filled in the part of their endowments to be destroyed. The envelopes containing the forms were then returned to the take authorities for their information. Meanwhile, responders were asked to indicate which take rate they had expected and how intensely they had experienced several emotions after having learned about the take rate. After completing the questionnaires and collecting all envelopes, subjects were privately paid outside the laboratory by the cashier who was not present during the experiment. Experimenters were not able to see what decisions subjects made in the game and how much they earned." Ibid.

- For both treatments, the intensity of these emotions is correlated with the take rate.

- "With effort, the probability of destruction…depends positively on the intensity of irritation and contempt. Without effort, the probability of destruction depends positively on the intensity of anger and contempt, and negatively on the intensity of happiness and joy."[80]

- Responders who destroy everything report more irritation than those who destroy only part.[81]

In these studies intensity of emotional reactions is a strong predictor of Responder behavior. And importantly, anger is by no means the only relevant emotion. Especially fascinating is that contempt is always present. This suggests a possibility that I have explored elsewhere: that prideful types — what Hobbes called glory-seekers — may play a critical role in upholding moral rules concerning fairness.[82]

In a recent study Fabio Galeotti has shown that the predictive value of emotional reactions can be considerably lessened if the Responders' destroy options are restricted to a fixed rate (2:1) for each unit taken.[83] Rather than Responders deciding how much to destroy in response to a taking, they simply opt to destroy at the fixed rate or not at all. In this treatment negative emotions remain correlated with the take rate, but have less predictive value of punishment. At low levels of punishment (for smaller takings) only contempt was of predictive value; at higher take rates (and so levels of punishment), those with higher levels of anger, irritation and contempt punished more, but this was significantly less predictive than under variable destruction rate treatments. Fixed rate punishment thus appears to blunt the predictive effect of emotions; it especially thwarts Responders' emotionally destroying their entire endowments in response to modest takings.

3.6 Expectations and Fairness

The suggestion, then, is that the mechanisms by which people uphold rules of justice

[80] Ibid., p. 420.

[81] Ibid., p. 417.

[82] "Why Being Touchy Protects Egoists from Exploitation: *Amour-propre* as a Basis of Fairness Norms" at www. gaus.biz

[83] Fabio Galeotti, "Do Negative Emotions Explain Punishment in Power-To-Take Game Experiments?" *Journal of Economic Psychology*, vol. 49 (2015): 1-14.

and fairness at considerable costs to themselves by no means depends on a magical sense of justice, Humean limited benevolence or even simply the internalization of social rules that supports our personal normative commitments. A plausible hypothesis is that emotional reactions, especially perhaps negative ones — such as guilt by perpetrators and anger, irritation and contempt by victims — are an important foundation of upholding rules of justice among strangers.[84] However, the mere fact that in Power-To-Take Games Responders' destructive behavior is significantly, in some cases powerfully, explained by their emotional reactions does not show that emotions are related to the rules of morality and fairness. However, other data does indicate a connection. In Bicchieri's important account of social norms, (roughly) a social norm is a behavioral rule *r* governing some type of behavior in a social network *S*, where most individuals in the social network prefer to conform to *r* on the conditions that (*i*) most others in *S* conform to *r* (an empirical expectation) and (*ii*) most people in *S* believe that most others in *S* ought to conform to it (a normative expectation).[85] Experimental evidence involving Dictator Games indicates that when normative and empirical expectations diverge, there is a strong tendency to align behavior with the empirical expectations.[86] An important finding in the Power-to-Take Games is that the Responders who punished very strongly tended to be (and in one study were exclusively) those who had expected lower take rates than they experienced (recall the presence of surprise).[87] This suggests that while negative emotions are well correlated with punishing behavior, this is strongly mediated by the punisher's *empirical* expectations about what others will do. However, as normative expectations have not been measured, we can only be tentative in suggesting that a norm is involved.

Thus far I have focused on Responders. Reuben and Winden studied the effect of Responders' punishment on Takers' take rate in a multi-stage Power-to-Take game.[88] They found that when Responders did not destroy, the Takers who increased their take rate in the second round tended to experience regret after the first round —

[84] That contempt is a significant emotion in almost all experiments suggests that pride is an important explanatory character trait.

[85] See *The Grammar of Society*, p. 11.

[86] Bicchieri and Erte Xiao, "Do the Right Thing: But Only if Others Do So."

[87] Bosman and van Winden, "Emotional Hazard in a Power-to-Take Experiment," p. 156; Bosman, Sutter and van Winden, "The Impact of Real Effort and Emotions in the Power-To-Take Game," p. 421; Galeotti, "Do Negative Emotions Explain Punishment in Power-To-Take Game Experiments?" p. 12.

[88] Reuben and Winden, "Fairness Perceptions and Prosocial Emotions in the Power-To-Take."

apparently regretting that they could have taken more and got away with it. Takers who did not experience destruction tended to increase their take rate in the second round; we might hypothesize that they were engaging in opportunistic behavior, and the absence of sanctioning encourages it. The behavior of Takers who did experience Responder destruction in the first round, however, was complex: some decreased their take rate while others did not. The key appears to be whether the Takers thought their taking was fair or unfair: those who took what they considered to be an unfair amount, to a significant degree reacted to Responders' punishment (i.e., destruction) by decreasing their takings. It is worth pointing out that in the first round these Takers apparently were willing to incur some guilt in return for high monetary gain *X*; in the second round they may have experienced an increase in guilt, which could well have led them to lower their taking.[89] However, Responder destruction did not have the effect of lowering the take rate of those Takers who thought their takings fair. This is consistent with other studies concluding that, in addition to the anger of punishers, effective punishment requires violators to experience guilt, say in recognition that they have violated their understanding of fairness or a social norm.[90]

I have considered experiments on Power-to-Take Games in some depth as they have focused on emotional reactions, and show that the typical fixation on anger misses a good deal of the relevant emotional reactions (and blinds us to the fascinating possibility that some reactions may be based on pride, rather than a form of moralistic aggression). We also should not make the false assumption that anger inherently leads to punishment. Experiments by Thulin and Bicchieri have shown that "moral outrage" — which is closely related to anger — underlies third-party *compensation* behavior, when norm violation has occurred. This is important: we should not suppose that negative emotions must be attached to a preference to punish violators, as opposed to compensating victims.[91]

[89] Ibid., p. 918. On the relation of guilt to interpersonal harm, see Mariëtte Berndsen, Joop van der Pligt, Bertjan Doosje and Antony Manstead, "Guilt and Regret: The Determining Role of Interpersonal And Intrapersonal Harm," *Cognition and Emotion*, vol. 18 (2004): 55-70.

[90] Hopfensitz and Reuben, "The Importance of Emotions for the Effectiveness of Social Punishment."

[91] "I'm So Angry I Could Help You: Moral Outrage as a Driver of Victim Compensation." It is important that Thulin and Bicchieri's target emotion appears distinctly moral; in one study emotions were measured, for example, on a 7-point scale from "Strongly Disagree" to "Strongly Agree" with statements such as "I feel angry when I learn about people suffering from unfairness" and "I think it's shameful when injustice is allowed to occur." These emotions are

4 SCALING-UP SOCIAL MORALITY

Given that we evolved in highly cooperative small group settings, it is hardly surprising that violation of social rules is associated with significant emotional reactions. The tale of Cephu the bad hunter is about attempted opportunistic cheating, group detection, deliberation, and the emotional storm that followed, albeit one that settled quickly once guilt had been admitted and punishment completed.[92] I have suggested that all this is far more than an ethnographer's vivid tale; the Reactive Emotions View helps explain not only the tale of Cephu, but has significant support in experimental evidence. I do not think it is much of a mystery either that we internalize moral rules and become devoted to them, or that our emotions are deeply involved in both moral judgment and action. The emotions seem especially important in inducing people to respond to defectors.[93]

One still might be tempted to resist: one might think this still is very much about small-group settings, so it may seem that we are back to a more sophisticated version of Bodo ethics.[94] However, recall that Ultimatum and Power-to-Take Games are one-shot anonymous interactions. They are games that make sense to people habituated to non-iterated rule-based interactions with strangers. Indeed, Ultimatum Games are played fairly similarly in all large-scale market-based societies. It is when we look at very small-scale societies that we can observe marked variation. The Machiguenga (of the Amazon Basin of southeastern Peru) for example, play the game in the originally expected "selfish" way, with many lower offers that are accepted. They also play public goods games with very high rates of defection.[95]

The type of moral guidance that I have sketched, with internalization of group rules, concern for the legitimacy of rules, and often strong emotional reactions at being treated in ways that defy our expectations, all scale-up to large-scale, anonymous interactions. It is, perhaps, just because in our original, and long habited, hunter-gather societies we developed this technology of social cooperation

---

moral emotions, presupposing a normative content, thus in my terms they appear to function as moral rules.

[92] The tale of Cephu seems to manifest both the "steaming up" and "cooling down" dynamics.

[93] See Bone, Silva and Raihani, "Defectors, Not Norm Violators, are Punished by Third-Parties."

[94] In some contexts Hardin intimates that the problem with all rule systems is that, because they depend on identification of a set of act-types, they cannot be usefully scaled-up to regulate dynamic societies with constantly changing act-types. Nichols and I analyze this idea in "Moral Learning in the Open Society."

[95] Henrich and Smith, "Comparative Evidence from Machiguenga, Mapuche, and American Populations."

that humans were able to so quickly and dramatically increase the scale of their societies at the beginning of the Holocene era. If the earliest societies really depended simply on rational self-interest regulated by self-interested punishment of defectors, it then *is* mysterious how humans could have left that small scale setting for huge cooperative orders so quickly.

The answer usually given by political philosophers, is, of course, "politics and the law." In large-scale societies, it is typically held, formal institutions, not the informal framework of social morality, do the work in securing cooperation. Now of course legal and political institutions are necessary for innumerable aspects of large-scale cooperation. No sane advocate of the importance of social morality or social norms would deny that. The question is whether these formal institutions supplant or supplement the basic framework of social rules and norms. Increasingly, I believe it is coming to be recognized that legal and political regulation without an underlying social normative framework is ineffective. Gerry Mackie has pointed out that there are hundreds of critical cases around the world in which practices — among them female genital cutting, caste discrimination, child marriage — have been widely criminalized yet continue to be practiced. Laws that depart from the basic moral and social norms of a society mostly likely will be ignored, often engendering contempt for the law. As Mackie, following Iris Marion Young, [96] concludes, "Criminalization is an appropriate response to a criminal injustice, a deviation from accepted norms, its harmful consequences intended, knowingly committed by identifiable individuals, whose wrongdoing should be punished. It is not an appropriate response to a structural injustice, in compliance with accepted norms, its harmful consequences unintended byproducts, and caused by everyone and no one. The proper remedy for a harmful social norm is organized social change, not fault, blame, punishment."[97]

In recent years students of social change have come to something of a consensus that effective legal regulation cannot stray too far from the underlying informal social rules.[98] One of the most striking "social experiments" based on this insight

[96] Iris Marion Young, *Responsibility for Justice* (Oxford: Oxford University Press, 2011).

[97] Gerry Mackie, "Effective Rule of Law Requires Construction of a Social Norm of Legal Obedience" in *Rethinking Cultural Agency: The Significance of Antanas Mockus*, edited by Carlo Tognato (Cambridge: Harvard University Press, 2015).

[98] In addition to ibid., see Bicchieri, *Norms in the Wild*; Bicchieri and Mercier, "Norm and Beliefs: How Change Occurs."

was that of Antanas Mockus, mayor of Bogotá in the late 1990s and early 2000s.[99] Mockus's aim was to harmonize legislation with social morality; he recognized that unless supported by the underlying informal moral and social framework, attempts to induce change though law would not succeed. For example, Bogotá was characterized by a very high rate of traffic fatalities in the mid-1990s, with widespread disregard for traffic regulations. Mockus distributed 350,000 "Thumbs Up/Thumbs Down" cards that drivers could display in response to dangerous driving by others, to drive home the message that such behavior was not only illegal, but violated the informal normative judgments of other drivers. Along with related programs, Bogotá witnessed a 63% decrease in traffic fatalities between 1995-2003. Similar programs based on harmonizing the law with informal social normative expectations led to decreases in water usage and, critically, homicides.

In lieu of an informal moral framework that coheres with the law, in a wide variety of cases (including traffic laws, which look like simply a coordination matter) we cannot expect the mass of citizens to conform unless coerced by high and effective penalties. And in the absence of such a framework we cannot expect those occupying positions in the formal institutions (in charge of administering those penalties) to be guided by its rules rather than taking the myriad opportunities for opportunistic enriching of themselves.[100] Institutions designed to promote cooperation can — and very often do — lead to kleptocracy.[101] Without the necessary foundation in an effective social morality, law and politics become simply additional devices by which some use power to extract from others.

5 CONCLUSION: COWS CAN BE COMPLEX

"[O]f all the differences between man and the lower animals," Darwin observes, "the moral sense or conscience is by far the most important….It is the most noble of all the attributes of man…Immanuel Kant exclaims, 'Duty! Wonderous thought, that worketh neither by fond insinuation, flattery, nor by any threat….'"[102] As Darwin

[99] For a short description of this experiment, see Antanas Mockus, "Building 'Citizenship Culture' in Bogotá" *Journal of International Affairs*, vol. 65 (Spring/Summer 2012): 143-46. For an in-depth treatment, see Mockus, "Bogotá's Capacity for Self-Transformation and Citizenship Building" in *Rethinking Cultural Agency*, edited by Tognato.

[100] See David Schwab and Elinor Ostrom, "The Vital Role of Norms and Rules in Maintaining Open Public and Private Economies" in *Moral Markets,* 204-227 at p. 209-11.

[101] See Friedman, *Moral and Markets,* chap. 5.

[102] Charles Darwin, *The Descent of Man*, second edition (New York: Penguin, 2004 [1879]), p. 120.

recognized, it is the invention of morality, self-control and conscience that allowed us to develop into one of the few eu-social species.[103] Darwin had no doubts that human morality and normative guidance was evolved, complex, and in many ways the defining feature of human social life.

Those whose work I most admire, in rightly seeking to avoid the sterility and unworldliness of so much moral and political philosophy, often turn to those models of clear-headed, empirically-informed, social philosophers: Hobbes and Hume. And I freely confess that it was the hard-headed beauty of *Leviathan* that hooked me on political philosophy. For others it was the empirically rich and moderate Hume that captivated them. But while many of us deeply we admire Hobbes and Hume, we must also acknowledge that their view of humans, and the ways they might solve their basic dilemmas of social life, were limited and too simple. In the last two decades we have discovered that humans are far more complex cooperators than we thought. Recognizing the importance of social-moral rules, their internalization and enforcement, is not an appeal to the mysterious but is required by attention to the facts.

Philosophy
University of Arizona

[103] Ibid., p. 133.