Punishment and Disagreement in The State of Nature

Jacob Barrett

1. The State of Nature

Hobbes famously declared that human beings living in the state of nature would find themselves engaged in a "war of every one against everyone" (*L* 3:8).[1] This state of universal war, he insisted, is the inevitable result of people "keeping company where there is no power to overawe them"—that is, of humans living together without a government to keep them in check (3:5). Thankfully, the situation is not so dire as it might seem for these poor denizens of the state of nature, as "reason suggesteth convenient articles of peace," "Laws of Nature" (3:14) or "*moral* laws" (26:37) whose widespread observance would conduce to everyone's interest and especially to their self-preservation. But while each may recognize the appeal of these rules, individuals cannot generally be trusted to comply without some "visible power to keep them in awe, and tie them by the fear of punishment to the… observation of th[e] laws of nature" (17:1). And, on Hobbes's view, only government can wield such mighty power.

Locke's analysis of the state of nature is similar.[2] He agrees that peace requires compliance with moral rules, and that humans in the state of nature would, through reason,

[1] All citations to Hobbes's *Leviathan* [L] and to Locke's *Second Treatise of Government* [ST] and *Essay Concerning Human Understanding* [EHU] are by chapter and paragraph number.

[2] But his definition of the term "state of nature" might not be: Locke often uses the term to refer to circumstances in which an illegitimate government is in place, and sometimes speaks of individuals being in the state of nature *in relation* to one another (Simmons 1989). Perhaps Hobbes does the same (Kavka 1986: 87-88). But, in any event, I always use the term "state of nature" narrowly, to refer to a circumstance in which humans live without government.

discover a "law of nature" directing them not "to harm another in his life, health, liberty, or possessions" (*ST* 2:6). He furthermore agrees that these rules must be enforced if they are to be obeyed: "the *law of nature* would… be in vain, if there were no body that in the state of nature had a *power to execute* that law" (2:7). But Locke rejects Hobbes's assumption that only government can provide enforcement. Individuals in the state of nature, too, have a "right to punish" one another for wrongdoing, and will exercise this right to deter such behavior (2:8). The trouble is that people will disagree about the content and application of the law of nature, and—since they must be "judges in their own cases"—these disagreements will lead to war (2:13). The state of nature therefore need not be a war of all against all, but may instead be a state in which conflicts intermittently erupt over moral disagreements, and in which people's "lives, liberties and estates" are therefore "very unsecure" (9:123). So, on Locke's view, government is not needed to provide enforcement power, so much as to consolidate this preexisting power behind an "established*, settled, known* law," a "known and indifferent judge," and a mechanism for giving its verdicts "due *execution*" (9:124-126).

There is a large literature on Hobbes's state of nature, much of which employs game-theoretic and other social scientific techniques to show why the strategic circumstances of the state of nature would indeed produce war (Gauthier 1986: ch. 2; Hampton 1986: chs. 2-3; Kavka 1986: ch. 3; Dodds and Shoemaker 2002; Vanderschraaf 2006; Chung 2015). But such analyses of Locke's state of nature are rare, and those few on offer omit the peace-enhancing role of punishment altogether, despite this being, on the interpretation I present, the key difference between Hobbes's and Locke's states of nature (Kavka 1986: ch. 3; Vanderschraaf 2006; Gaus 2018; Kogelmann and Ogden 2018; Bruner 2018). In this paper, I fill this lacuna in the literature by providing accounts of Hobbes's and Locke's states of nature along the lines I have just sketched, which emphasize both the role that punishment

plays in transforming a state that would otherwise be a Hobbesian universal war into Locke's state of "peace, good will, mutual assistance and preservation," and the role that moral disagreement plays in rendering this peace insecure (*ST* 3:19). Throughout, my aim is not purely interpretive, but also to demonstrate the plausibility of Locke's analysis by drawing on classical and experimental game theory, and more generally to bring the philosophical literature on the state of nature into contact with the economics literature on the conditions under which punishment can stabilize peaceful cooperation.

I begin with a simple game-theoretic analysis of the dynamic of conflict that characterizes Hobbes's state of nature which draws out the essential features of the (typically much more complicated) models in the literature, before turning to a discussion of why mere recognition of the laws of nature cannot change this dynamic. The reason, I argue, is that the situation can be illuminatingly modeled by what experimental economists call a "public goods game," the typical results of which exhibit precisely the same dynamic as our initial analysis predicts. I then provide an interpretation of Locke's state of nature that modifies the Hobbesian state by attributing to individuals the ability and willingness to punish wrongdoing, and draw on experimental (and computational) results concerning the effect of punishment in public goods games to show that while such collective enforcement may plausibly stabilize peaceful cooperation under conditions of moral agreement, it cannot do so under conditions of disagreement. Thus, on the Lockean story I tell, the reason we need government is not to keep a few wicked or relentlessly self-interested individuals in line—as our discussion of Hobbes's state of nature will suggest—but rather to maintain peace among those who disagree about morality. I close by briefly comparing Hobbes's authoritarian solution to the problem of moral disagreement to Locke's liberal solution.

2. The War of All Against All

On Hobbes's account of human nature, humans pursue the objects of their desires. But since we never succeed in fulfilling all of our desires, our concern is not only to satisfy our current desires, but also to achieve *felicity*—"continual success in obtaining those things which [we] from time to time desire" (*L* 6:58)—and therefore to "assure the power and means to live well" going forward (11:2). So denizens of the state of nature aim "principally" at self-preservation (which is, after all, a precondition for the achievement of any future good), but also at power and the attainment of particular things (13:3). And since they are roughly equal in strength, intelligence, and ability, whenever two desire the same object, both have an "equal hope in the attaining" of it, and so find themselves in "competition" for it (13:3). This competition "maketh men invade for Gain" (13:7), and the recognition that others invade for gain also leads to anticipatory attacks out of "diffidence" that, otherwise, others will attack one first (13:7). A third cause of conflict is "glory," or the pleasure some take in their own "power and ability" (6:39). Glory amplifies the tendency to invade for gain, which in turn amplifies the tendency to attack out of diffidence.

Hobbes's analysis of conflict has, in recent years, been helpfully modeled with certain tools from game theory. These models standardly assume that when two individuals, A and B, meet in the state of nature, they have two options: they may attack or refrain. If we say that for A to "exploit" B is for A to gain at B's expense—as occurs, say, when A seizes B's resources, or kills or wounds B thus removing or weakening a competitor—then this yields four possibilities, as displayed in the following matrix (Vanderschraaf 2006: 249):

		B: Refrain	B: Attack
A	Refrain	Peace	B exploits A
A	Attack	A exploits B	Battle

Figure 1.1: The State of Nature Game

Such models then try to show that though each would prefer peace to battle, the latter is the inevitable result of agents pursuing their preferred options. And thus, they are led to war.

Although early analyses of Hobbes's state of nature assumed that individuals all share the same preferences (Gauthier 1986: ch. 2, Hampton 1986: ch. 3), more recent ones emphasize that Hobbes allows for two different types, and I will closely follow these latter analyses myself (Kavka 1986: ch. 3; Dodds and Shoemaker 2002; Vanderschraaf 2006; Chung 2015). Consider the following key passage:

> because there be *some*, that taking pleasure in contemplating their own power in the acts of conquest, which they pursue farther than their security requires; if *others*, that otherwise would be glad to be at ease within modest bounds, should not by invasion increase their power, they would not be able, long time, by standing only on their defence, to subsist (*L* 13:4, emphasis mine)

Here, Hobbes distinguishes *dominators*, who prefer exploiting others to peace, from *moderates*, who prefer peace to exploiting others (Kavka 1986: 97). Although both dominators and moderates have some motivation to exploit others to achieve short-term gain, moderates have competing motivations that outweigh this, and so have an all-things-considered preference for peace over exploitation. On a pessimistic reading of Hobbes's psychology, this is because moderates recognize that peaceful interaction is in their long-term self-interest; on an optimistic reading, it is because they are "good natured," taking pleasure in

others doing well (*L* 6:2), and perhaps even "car[ing] deeply about justifying themselves to others" (Lloyd 2009: 89). Dominators, however, have an all-things-considered preference for exploiting others over peace because they lack good nature and are "short-sighted" in the sense that they fail either to recognize or to sufficiently care about the long-term benefits of peaceful cooperation (Hampton 1986: ch. 3). This is in part because they are glorious, seeking power for its own sake even beyond the long-term benefits it brings (Chung 2015).

Besides this, dominators and moderates share the same preferences: for peace over battle, and for battle over being exploited. So if A is a dominator, her preference ordering is: A exploits B, peace, battle, B exploits A. And if A is moderate, her ordering is: peace, A exploits B, battle, B exploits A. There are thus three specifications of the state of nature game to consider, depending on whether two dominators meet, two moderates meet, or a dominator meets a moderate. If the first occurs, we have a *prisoner's dilemma*:[3]

		Dominator B: Refrain	Dominator B: Attack
Dominator A	Refrain	2, 2	4, 1
	Attack	1, 4	3, 3

Figure 1.2: The State of Nature Game (Prisoner's Dilemma)

Here, though each prefers peace, dominance reasoning leads them to battle. Each thinks to herself: if the other refrains, I get my 1st choice rather than my 2nd, and if the other attacks, I

[3] In the following matrix, "4, 1" ("x, y") represents that the row player A gets her 4th (xth) choice, and the column player B gets her 1st (yth) choice. I adopt this convention throughout.

get my 3rd choice rather than my 4th, so I do better to attack no matter what, and I should attack. And thus, whenever two dominators meet, they fight.[4]

Suppose, second, that two moderates meet. Then they face an *assurance game*:

		Moderate B: Refrain	Moderate B: Attack
Moderate A	Refrain	1, 1	4, 2
	Attack	2, 4	3, 3

Figure 1.3: The State of Nature Game (Assurance Game)

This time, each does better to refrain if the other refrains, but better to attack if the other attacks. So both peace and battle are live possibilities (both are Nash equilibria).[5] If A can be *assured* that B will refrain, A will refrain herself, since A most prefers peace. But if A predicts that B will attack, A will attack, too, since A prefers battle to being exploited. The outcome

[4] A standard objection here is that I have assumed dominators are playing "one-shot" as opposed to "iterated" games in which they repeatedly interact with the same partners (Hampton 1986: ch. 2; Kavka 1986: ch. 3), and that, in iterated games, even individuals with the preferences of dominators in one-shot games may achieve peace by adopting a strategy like "tit-for-tat" (Axelrod 2006). This objection fails because our description of dominators already takes into account the prospects of repeated interaction: dominators are defined as those who are sufficiently short-sighted and unmoved by good nature that, when they encounter others, they prefer exploiting them to peace despite the long-term benefits of repeated peaceful interaction. So iterating the game makes no difference to the behavior of dominators, or, therefore, to our analysis.

[5] A Nash equilibrium obtains when no player can gain by unilaterally changing her strategy.

therefore depends on whether each is assured the other will refrain—more on which shortly.

Finally, if a moderate, A, meets a dominator, B, we have an *assurance dilemma*:

		Dominator B: Refrain	Dominator B: Attack
Moderate A	Refrain	1, 2	4, 1
	Attack	2, 4	3, 3

Figure 1.4: The State of Nature Game (Assurance Dilemma)

Here, B employs the same reasoning as in the prisoner's dilemma: she does better to attack no matter what, so she will attack. And A, recognizing this, will also attack, since she prefers to attack when attacked. So, whenever a dominator meets a moderate, they battle.

So far, our analysis suggests that though dominators always attack, moderates may achieve pockets of peace in cases where each trusts the other to refrain. But once we recognize that moderates in the state of nature are often uncertain about whether they are interacting with moderates or dominators, conflict ramps up between moderates as well. To be clear, we need not oversell the degree of uncertainty here: friends and family members may identify each other as moderates, so there may be room for small "confederacies" in the state of nature (*L* 15:5). Rather, the problem arises when a moderate A meets a stranger B, since, in this case, A lacks assurance that B will refrain: not only may B be a dominator who always attacks, but B may even be a moderate who will attack because she lacks assurance that A will refrain (Kavka 1986: 104; Dodds and Shoemaker 2002). More carefully, let us say that a moderate will refrain from attacking a stranger only if she assigns sufficiently high probability p to strangers refraining. And let us say that a moderate is *trusting* if this threshold is met, and *diffident* if it is not. Although the threshold value of p will vary depending on the

relative intensity of A's preferences—a variable omitted in the ordinal matrices above—it seems safe to assume that this value will be fairly high for most moderates, given the high cost of being exploited. And this threshold is unlikely to be met in the state of nature, where distrust runs rampant (Vanderschraaf 2006; Chung 2015).

To see why, note that since an individual's evaluation of p concerns the behavior of strangers, her estimation of the probability p that her partner will attack must correspond to her estimation of the *frequency* of attackers in the population, such that she will form this estimate on the basis of her past experience: the more she witnesses others attacking, the lower her estimation of p will be, and the more she witnesses others refraining, the higher her estimation will be. And so, to summarize the lesson of certain recent (though highly technical) models of Hobbes's state of nature, even a few dominators in the state of nature may produce universal war (Vanderschraaf 2006; Chung 2015). For even if we assume that all moderates start out trusting, dominators will attack, and moderates—beginning with those with the highest threshold p values—who survive or witness these attacks will become diffident. As time goes on, both dominators and these now diffident moderates will attack, and as trusting moderates witness these attacks, those with the highest threshold p values will also become diffident, and so they will begin to attack, and so on—until, eventually, the entire population is full of dominators and diffident moderates. Distrust, in other words, is contagious, and a small contingent of dominators can spread it throughout a large population, instigating and maintaining war. For "the nature of war consisteth not in actual fighting but in the known disposition thereto during all the time there is no assurance to the contrary" (*L* 13:8). And this assurance cannot be had in the state of nature.

3. Convenient Articles of Peace

After outlining some proto-version of this argument, Hobbes changes his tune. There are reasons to think, he suggests, that people may achieve peace after all. For all may recognize, through reason, precisely the situation they are in, and therefore that they would be better off under peace than war:

> The passions that incline men to peace are fear of death, desire of such things as are necessary to commodious living, and a hope by their industry to obtain them. And reason suggesteth convenient articles of peace, which men may be drawn to agreements. These articles…. are called the Laws of Nature (*L* 13:14).

"A law of nature," Hobbes explains, "is a precept or general rule, found out by reason, by which a man is forbidden to do that which is destructive of his life" (14:3)—the "science" of which is "the true and only moral philosophy" (15:40). The fundamental law of nature is "*to seek peace and follow it*" (14:4), and all can be summed up under the motto, "*Do not that unto another, which thou wouldst not have done to thyself*" (15:35). In the state of nature game, the laws of nature therefore require one to refrain rather than to attack. Their widespread compliance suffices to establish peace.

Unfortunately, the mere recognition of these laws cannot secure their compliance. Dominators still prefer to violate the rules and attack, and moderates still prefer to attack unless they are assured their partners will refrain. The laws of nature "oblige *in foro interno*, that is to say, they bind to a desire that they should take place; but *in foro externo,* that is, to the putting them in act, not always" (15:36). For though everyone in the state of nature should want the laws of nature to "take place"—to obtain as a behavioral regularity—someone who complies without "sufficient security" that others will "makes himself a prey to others, and procures his own certain ruin, contrary to the ground of all laws of nature"

(15:36). To establish peace, the laws of nature must therefore be not only recognized, but enforced. For "the laws of nature, of themselves, without the terror of some power to cause them to be observed, are contrary to [the] natural passions" that drove our above analysis (17:2).

To help understand Hobbes's reasoning here, note that widespread compliance with such laws is a *public good*. If all comply this is better for everyone, but since individuals have a strong incentive to defect, many "free-ride" and the public good is undersupplied. Indeed, since violations of the laws of nature sow distrust and are therefore contagious, violating the rules at least typically harms not only the individual that is attacked, but also the public at large: attacking decreases moderates' trust and therefore increases the probability of them attacking in the future. And this is bad for everyone—for dominators and moderates alike.

One useful way to model this situation is therefore to note that when individuals play the dyadic (one-on-one) state of nature game they are simultaneously playing a population-wide public goods game. In this game, which is a staple of experimental economics, some number of individuals must choose either to cooperate by adding some money to a common pot or defect by keeping it for themselves. The money in the pot is then multiplied by some factor, and distributed equally among the players. This models a public good because though each benefits from others cooperating and donating, each achieves a greater personal benefit by defecting and not donating: for example, for a population size of ten, a contribution value of $10, and a multiplier of five, donating gives everyone else $5 (the $10 contribution is multiplied by five and split among ten individuals), but results in a net loss of $5 to the donator (donators contribute $10 and receive only $5 of it back). Thus, if this game is played a single time, and if we assume people aim to maximize their monetary earnings, we have a multi-person prisoner's dilemma. Each individual will defect and retain $10, even though

each would get $50 if everyone cooperated (Ledyard 1995).[6]

In the state of nature, to cooperate is not, of course, to add money to a pot, but to refrain from violating the laws of nature when one can gain resources or power from violating them and attacking. And the reason this benefits everyone is not because it provides them with money, but, again, because it increases rather than decreases trust. Yet the game remains structurally parallel to the state of nature, and it therefore provides an illuminating model of it—in no small part because the results of experiments involving such games appear to be driven largely by the interactions of two familiar types. Some are *free-riders* who care only about the monetary gains of defection and so never contribute to the pot. They act just like dominators who violate the laws of nature whenever they gain by doing so. But the empirical evidence overwhelmingly suggests that others are *conditional cooperators* who are willing to forego this gain and contribute on the condition that they expect enough others do the same (Fischbacher et al. 2001; Chaudhuri 2011: 51-56). They act just like moderates who comply with the laws of nature only when they have "sufficient security that others shall observe the same laws" (*L* 15:36). Of course, the underlying motivations of experimental participants and Hobbesian agents fighting for survival clearly differ, but what matters here is that their behavioral profiles are the same: both free-riders and dominators always defect, while both conditional cooperators and moderates cooperate only when they expect a sufficient number of others to do so as well. The results of experiments involving free-riders and conditional cooperators in public goods games may therefore shed light on the behavior of dominators and moderates in the state of nature.

In the last section, we described a phenomenon of contagion whereby a few

[6] In many games of this sort, including most I go on to cite, individuals must choose not only whether, but how much, to contribute. I omit this detail to simplify discussion.

dominators can lead a mixed group, even one predominated by moderates, to universal war. And remarkably enough, this same dynamic typically plays out when public goods games are iterated—played repeatedly by the same participants over a number of rounds—in the economics lab (Ledyard 1995; Chaudhuri 2011). In the first round, some contribute to the pot: these are conditional cooperators, analogous to moderates who expect enough others to comply. But a sizable group don't: these are free-riders or diffident moderates, analogous to dominators or conditional cooperators who don't expect enough others to comply. In each subsequent round, individuals are made aware of contribution patterns in the last round, and fewer contribute: this is explained by conditional cooperators or moderates updating their beliefs about how likely others are to contribute or comply in response to information about past defections. And by the final round, hardly any one cooperates: war prevails.

Thus we find, playing out in real time, the dynamic of conflict that permeates Hobbes's state of nature. Few expect the laws of nature to be widely complied with, so few are willing to comply with them. And those moderates who do comply given their belief that others will do so quickly realize their mistake, and cease to comply "for caution against all other men" (*L* 17:2). To ensure peace, individuals must therefore establish some enforcement mechanism that uses punishment to transform the state of nature game into an enforcement game in which everyone, factoring in the likelihood and severity of being punished if they attack, prefers to refrain than to attack no matter what their partner does:[7]

[7] In the following matrix, I assume, for illustration's sake, that each prefers mutual refraining to refraining while the other attacks, and attacking while the other refrains to mutual attacking. But all that matters is that each prefers both outcomes in which she refrains to either in which she attacks.

		B: Refrain	B: Attack (Punished)
A	Refrain	1, 1	2, 3
A	Attack (Punished)	3, 2	4, 4

Figure 2.1: The Enforcement Game

This, on Hobbes's view, is why we need government. For only government can ensure that dominators gain more from refraining, and that moderates have the assurance they need that others will refrain in order to refrain themselves.

4. The Right to Punish

Although Locke's analysis of the state of nature differs in important respects from Hobbes's, he agrees, as I have suggested, on two crucial points. First, Locke agrees that widespread compliance with moral rules—which, on his view, are also discoverable through reason, and proclaim that "no one ought to harm another in his life, health, liberty or possessions"—is needed to avoid conflict in the state of nature (*ST* 2:6). The law of nature provides "peace and safety" and "mutual security"; it "willeth the peace and preservation of all mankind" (2:7). Second, though Locke believes that people are morally motivated—due either to the expectation of "[t]he Rewards and Punishments of Another Life" (*EHU* 21:70; Colman 1983: 72-74), or to a special pleasure we take in "doing our duty" (Locke 1997: 319; Sheridan 2007)—he agrees that such motivation is not enough to command widespread compliance in the absence of enforcement. Again: "the *law of nature* would… be in vain, if there were no body that in the state of nature had a *power to execute* that law, and thereby preserve the innocent and restrain offenders" (*ST* 2:7).

Indeed, though Locke and his commentators are less explicit on this point, it is clear

that the same analysis we have provided of Hobbes's state of nature maps neatly onto Locke's, so long as we continue to assume that moral rules go unenforced. For Locke allows that both dominators and moderates exist in the state of nature. Some are "degenerates" (*ST* 2:10) who are unmotivated by morality, "have no other rule, but that of force and violence," and so attack whenever this yields a short-term gain (3:16). These are Hobbesian dominators (Ashcraft 1968: 904). Others sometimes refrain from attacking out of moral motivation, but attack when another has attacked first, or when they believe they "have discovered an enmity to his being"—that is, when they fear that the other will attack them if they don't anticipatorily attack first (*ST* 3:16). They therefore decide whether to attack based on their estimation of how likely others are to attack, and so, though their underlying psychologies may differ, act just like moderates (Vanderschraaf 2006: 250-251). So both dominators and moderates populate Locke's state of nature, and for reasons explored in the last two sections, this is enough to generate war in the absence of enforcement. In fact, the problem is even worse in the Lockean case, since Locke maintains that moral motivation can "curb" but not override "exorbitant Desires" (*EHU* 3:13) such that "the greater part [are] no *strict* observers of equity and justice" (*ST* 64). In other words, not only do dominators and diffident moderates always attack, but even trusting moderates attack in the face of great temptation, and this renders enforcement all the more necessary (Colman 1983: 183-185).[8]

[8] One might protest that Lockean moderates will not anticipatorily attack, since Locke holds, against Hobbes, that war must be "declare[d] by word or action" rather than assumed from a lack of assurance (*ST* 3:16). Yet despite what some have thought, it does not follow that Lockean moderates will refrain from anticipatorily attacking "in the absence of all specific evidence about [others' intentions]" (Kavka 1986: 90). For even if this is wrong, self-interest will typically outweigh moral motivation in cases where moderates fear that, if they do not

So both Locke and Hobbes believe that widespread compliance with moral rules is necessary if people are to live in peace, but that compliance requires enforcement. From this, Hobbes concludes that these rules will not be complied with in the state of nature, since there is no government to enforce them. But Locke's insight is that there is another option. People in the state of nature, too, can provide enforcement, since "the *execution* of the law of nature is, in that state, put into every man's hands, whereby every one has a right to punish the transgressors of that law to such a degree, as may hinder its violation" (*ST* 2:7). This right to punish is not only a right of the "injured party" to retaliate, but a right of anyone to punish violators of the law of nature: though only the "injured party" has the right "of taking reparation" from her attacker, "the right of punishing is in everybody" (2:11). Here, I set aside reparation and focus on punishment. The question is whether people can collectively enforce the law of nature and transform the state of nature game into the enforcement game, or whether Hobbes is right that only government is up to the task.

One initial barrier to collective enforcement concerns the difference between the existence of a right and its exercise. That people in the state of nature have a right to punish does not imply that they will punish, and if no one is sufficiently motivated to do so, then this right will lie inert.[9] Now, on Locke's account, one reason individuals may actually punish is that some are motivated by *anger*: an "uneasiness or discomposure of the mind, upon the receipt of any injury, with a present purpose of revenge" (*EHU* 20:12). Such individuals may

---

anticipate, others will attack first. And Locke denies that this is wrong anyway, since the law of nature only binds "when [one's] own preservation comes not in competition" (*ST* 2:6).

[9] Indeed, Hobbes himself held that individuals in the state of nature have a right to punish (*L* 28:2), but he apparently did not believe that this right would be exercised often enough to change the dynamics of the state of nature.

therefore retaliate against those who have wronged them out of anger, and perhaps also to enhance their reputation: if they become known as retaliators, this may deter others from attacking them in the future. Unfortunately, such retaliation is, on it own, unlikely to stabilize peace. For punishment to serve its deterrent function, it must be reliable and severe enough that dominators predict the expected cost of punishment to outweigh the expected benefit of wrongdoing. But if individuals only need fear retaliation, punishment will not meet this requirement: an individual's "own single strength, hath not force enough to defend himself from injuries, or to punish delinquents" (*ST* 11:136). Thankfully, Locke has two resources to explain why individuals will also be motivated to engage in third-party punishment, so that enforcement need not depend on retaliation alone.

The first motivation to engage in third-party punishment involves the long-term benefit that the deterrent effect of punishment provides to the punisher:

> In transgressing the law of nature, the offender… becomes dangerous to mankind… Which being a trespass against the whole species, and the peace and safety of it, provided by the law of nature, every man… may bring such evil on any one, who hath transgresses that law, as may make him repent the doing of it, and thereby deter him, and by his example, others, from doing the like mischief (*ST* 2:8)

Here, Locke suggests that punishing offenders both specifically deters recidivism, and generally deters others from wrongdoing. Individuals will therefore be motivated to punish offenders in order to deter future wrongdoing, thereby decreasing the likelihood that others will injure the punisher, either directly through attacking them, or indirectly by decreasing trust and increasing rates of attack overall.

Locke's other resource is moral motivation. The law of nature imposes a fundamental duty to "*preserve the rest of mankind*" (*ST* 2:9), and since punishment tends to this

preservation via deterrence, individuals have a derivative duty to punish wrongdoers (Simmons 1991: 325-326). Indeed, on Locke's account, the moral justification of punishment depends precisely on this deterrent effect: "each transgression may be *punished* to that *degree,* and with so much *severity,* as will suffice to make it an ill bargain to the offender, give him cause to repent, and terrify others from doing the like" (*ST* 2:12). Of course, we should not assume that moral motivation to punish is overriding. Instead, Locke's psychology implies that moderates are to some extent morally motivated to punish wrongdoing, that they are more motivated in cases where they expect the deterrent effect of this punishment to benefit them, and that they will therefore punish when these two motivations combine to outweigh their motivation to avoid the cost of punishing. If this cost is low, punishment may therefore stabilize peaceful cooperation even in the absence of government. For as long as enough are willing to engage in punishment that the threat they collectively pose makes individuals prefer to comply than to defect, the state of nature game will transform into the enforcement game, and peace will obtain.

To examine this possibility, let us consider a modification of the public goods game: a public goods game with punishment. Here, as before, individuals must choose whether to donate money to a pot (analogous to complying with the laws of nature when one can personally benefit by violating them) or keep it for themselves (analogous to violating them and gaining that benefit), where contributions are then split among its participants (analogous to compliance benefiting everyone by increasing rather than decreasing trust). But after this initial phase, others become aware of who cooperated and who defected, and are then given an opportunity to pay some cost to punish defectors: say, by paying $5 to remove $10 from the defector (analogous to punishing wrongdoers in the state of nature at some cost). This two-phase game is then iterated a number of times, the hope being that

individuals will punish those who defect, and that this will deter future defection and ultimately stabilize peaceful cooperation among the participants (Fehr and Gächter 2000; see also Ostrom et al. 1992).[10]

And sure enough, though we again need not assume that the underlying psychology of experimental participants is the same as Lockean agents in the state of nature, we do find punishment behaviorally manifesting in laboratory settings in exactly the way Locke's psychology predicts. As mentioned above, in an iterated public goods game without punishment, the typical result is that some contribute in the first round, but rates of contribution rapidly decay. But in the typical iterated public good games with punishment, closer to the opposite occurs. In the first round, some contribute to the pot and others don't, after which some subset of the population punish those who failed to contribute. In each subsequent round, more people contribute for fear of punishment, until by the end, the vast majority does the same (Fehr and Gächter 2000). In line with the behavioral predictions of Locke's psychology, such punishment is especially common when individuals engage with the same partners each round, since they may expect to benefit in future rounds from their punishment's deterrent effect. But even in cases where individuals are randomly matched with new partners each round, such that only moral motivation could explain such "altruistic punishment," a significant amount of punishment still occurs—enough to stabilize a higher rate of contribution than the game with no punishment, though not as much as the game where both self-interested and moral motivation operate (Fehr and Gächter 2002). And since, in the state of nature, both motives are frequently operative, we should expect the state of nature to more closely resemble the former, and for there therefore to be widespread

[10] In such games, individuals must often choose not only whether, but how severely, to punish others. I again omit this detail to simplify our discussion, but see fn. 12 below.

enforcement and compliance with the laws of nature. On our Lockean analysis, all this requires is that the cost of punishment remains low. And, indeed, this same requirement appears to hold empirically in the lab (though, as we will see in the next section, it is not the only relevant requirement): punishment is effective at stabilizing cooperation only when its cost is low both absolutely and in relation to its deterrent power (Egan and Riedl 2008).

Thus, on this first take on Locke's state of nature, the state of nature is best modeled as a public goods game with punishment. The law of nature can be enforced without government, and, contra Hobbes, "*the state of nature and the state of war*… are as far distant, as a state of peace, good will, mutual assistance and preservation, and a state of enmity, malice, violence, and mutual destruction are one from another" (*ST* 3:19). But Locke is no anarchist: he does not really view the state of nature with such rose-colored glasses. Mere paragraphs later, he acknowledges that it may involve a breakdown of peace, and that "this is one great reason of men's putting themselves into society, and quitting the state of nature" (3:21). Indeed, Locke asserts that the state of nature "is full of fears and continual dangers" (9:123), "mutual grievances, injuries and wrongs," and "strifes and troubles" (7:91). Some see this as a "central contradiction" in Locke's theory (Macpherson 1962: 241). But, more plausibly, Locke believed the state of nature could produce a range of outcomes, from more to less peaceful, but that it would always be a state in which peace, even when achieved, is insecure (Simmons 1989: 458-459). So let us turn to the causes of this insecurity now.

5. Confusion and Disorder

The analysis of the last section implicitly assumed that people agree about the content and application of the law of nature, and that people will accept punishment without resistance: in the experiments cited, there is no room for disagreement about whether people have

contributed or complied in the last round, and retaliating against punishment is simply not an option. Problems begin to emerge, however, once we realize that neither of these assumptions is warranted in the state of nature. People may disagree about the content or application of the law of nature: "though the law of nature be plain and intelligible to all rational creatures; yet men [are] biassed by their interest, as well as ignorant for want of study of it" (*ST* 9:124). And they may resist punishment, retaliating against those who punish them: "[t]hey who by any injustice offended, will seldom fail, where they are able, by force to make good their injustice" (9:126). Relying on private enforcement therefore leads to "nothing but confusion and disorder" (2:13). And thus, we must give up our right to punish, and establish a government that unifies this enforcement power behind an "*established*, settled, known *law"* and "*a known and indifferent judge*" to interpret and apply it (9:124-126).

Let us begin by continuing to assume that individuals agree about the content and application of morality, so that we can examine why social breakdown might occur even without disagreement. Again, the public goods game helps us to appreciate the issue. For, despite optimistic interpretations of early results from public goods games with punishment, more recent experiments have called this optimism into question (Chaudhuri 2011: 56-69). The basic problem arises when we allow individuals not only to *prosocially* punish those who defect, but also to *antisocially* punish those who comply (Hermann et al. 2008). Now, initially it might seem puzzling why individuals in the state of nature would engage in antisocial punishment given our provisional assumption of moral agreement: everyone typically benefits from others complying with the rules, so individuals rarely have an incentive to punish compliance. This puzzle evaporates once we recognize that there is an important case where, even under conditions of agreement, antisocial *counter*-punishment provides the punisher with a significant benefit, which dominators, at least, are likely to pursue: if a

dominator retaliates against someone who has prosocially punished him, he may deter her and others from punishing him in the future. As Locke recognized, this sort of counter-punishment drives up the cost of prosocial punishment and is thus highly destabilizing: "such resistance many times makes the punishment dangerous, and frequently destructive, to those who attempt it" (*ST* 9:126). For if there is a significant risk that my attempt to enforce the law of nature will be resisted by a wrongdoer, I may be unwilling to bear this cost. And if enough share this unwillingness, there will no longer be sufficient enforcement of the law of nature to deter dominators, trust will spiral downwards, and war will ensue.

This dynamic has been observed in experimental settings in which the iterated public goods game with punishment is modified so that individuals also have the opportunity to counter-punish those who punished them last round. These opportunities are often taken, fewer become willing to punish defections for fear of such counter-punishment, and cooperation breaks down (Nikiforakis 2008). Interestingly, however, if we tweak the experimental setup, counter-punishment opportunities can have a neutral or even positive effect on cooperation (Cinyagabuma et al. 2006). This is because counter-punishment can play two distinct roles. First, individuals may use counter-punishment to antisocially retaliate against prosocial punishment, thus increasing the cost of prosocial punishment and destabilizing cooperation. But, second, they may employ counter-punishment as a prosocial response to antisocial (counter-)punishment, thus deterring antisocial punishment and helping to stabilize cooperation. If individuals in experimental settings only have the former option, cooperation breaks down; if they only have the latter, cooperation sustains (Denant-Boemont et al. 2007). In the state of nature, however, we should expect individuals to engage in both forms of punishment. So it remains an open question whether the antisocial or prosocial punishers will carry the day, and whether war or peace will result.

As a first step in answering this question, note that whereas dominators in the state of nature may be strongly motivated to antisocially *retaliate* against those who prosocially punish them, they have little incentive to engage in *third-party* counter-punishment in defense of other wrongdoers. Widespread observance of the law of nature is a public good, so while dominators may wish to deter others from enforcing the law of nature against them, it is typically to their benefit that it is enforced against others. This is a fundamental asymmetry between antisocial and prosocial punishment, because moderates have deterrence-based and moral motivations to engage in prosocial third-party punishment as well. Of course, moderates may be even more motivated to engage in retaliation thanks to their extra anger and reputation-based motivations to do so. And perhaps these extra motivations will lead some to prosocially retaliate even given a high threat of antisocial counter-punishment: if I am wronged by someone, I might be moved to retaliate by anger or by the worry that if I don't, I will become known as someone who is weak, and so easily "preyed" upon. But once again, such prosocial retaliation is unlikely to stabilize peace, which requires enough moderates to engage in third-party punishment that enforcement is sufficiently reliable and severe. The real question, then, is whether those moderates who are willing to engage in third-party punishment when the cost is sufficiently low can keep these costs down by providing an effective deterrent to those who would otherwise escalate these costs through antisocial retaliation. Only then will they provide reliable enforcement of the laws of nature.

Although the possibility has not yet been studied in the economics lab, there are good theoretical and anthropological grounds for thinking that moderates may indeed accomplish this feat. The key is to realize that, unlike the experiments we have been considering in which punishment is *uncoordinated* and carried out by individuals, in the real world, punishment is often *coordinated* and carried out by groups (Boyd et al. 2010)—a fact

that Locke appears to recognize when he notes that a violator of the law of nature "renders himself liable to be destroyed by the injured person, *and the rest of mankind, that will join with him in the execution of justice*" (ST 15:172, my emphasis), that is, by anyone who "joins with [the injured party] in his defence, and espouses his quarrel" (ST 3:16). Such coordination allows for a significant decrease in the cost of punishment, because the larger the group, the lower the cost of punishment for each. There is "strength in numbers," since if a group gangs up on a wrongdoer, he will be less able and willing to put up resistance or to antisocially retaliate afterwards knowing that he is so greatly outnumbered (Boyd et al. 2010: 149).

Models of coordinated punishment see it as occurring in two steps. First, individuals signal their willingness to punish wrongdoing on the condition that enough others signal the same; second, if this quorum is met, the group gets together and punishes wrongdoing when it occurs. And computational models confirm that if individuals can coordinate prosocial punishment in this way, this can indeed sustain cooperation for a wide range of plausible parameters concerning the rate at which the cost of punishment declines with larger groups, the number of prosocial punishers in the population, and the cost such punishers are willing to bear (Boyd et. al. 2010).[11] Now, such models remain limited insofar as they do not also permit antisocial punishers to coordinate their punishment. But this is no defect in the current context, since this is exactly the asymmetry we have noted. Moderates' shared

[11] A complication arises if there are "liars" who signal that they are willing to punish but fail to follow through. Boyd et al. eliminate this possibility by stipulating that signaling is costly, and that "[t]he cost of signaling [one's intention to punish] is high enough so that it does not pay to signal and then fail to punish" (618). Actually, we need only claim that the cost of signaling *and then failing to follow through* is high enough that few will do it, as will occur when there are enough sincere punishers that liars expect to be punished for their treachery.

morality and willingness to engage in third-party enforcement allows them all to take the same side against dominators who would perhaps be willing to antisocially retaliate against lone prosocial punishers but who know better than to pick a fight with the group (DeScioli 2016). But dominators cannot similarly coordinate because they are generally motivated only to engage in antisocial retaliation, and so will not typically join up with other wrongdoers when the moral mob comes after them.

Under conditions of moral agreement, moderates should therefore be able to keep the cost of prosocial punishment relatively low through coordinating their punishment, thus maintaining peace through collectively enforcing the law of nature. And so, the real trouble in Locke's state of nature can only arise when we relax the assumption of moral agreement and admit, with Locke, that people will disagree over both the content and application of morality (Nozick 1974: 11-12; Parry 1978: 59; Gaus 2018; Bruner 2018). For, in this case, what one sees as rightful compliance with the laws of nature others may see as wrongdoing, and such disagreements may not only prevent individuals from coordinating their punishment of dominators, but may even lead to war among morally motivated moderates. Consider the following example. I believe you have broken the law of nature by stealing my food, so I retaliate by burning down your house. In my eyes, this punishment is prosocial and admirable, but you disagree. You think that burning down your house was antisocial and wrong: that I "miscite, or misapply" the law of nature in punishing you in this way (*ST* 9:136). You thus retaliate and punish me back, and I similarly respond with counter-retaliation.[12] We are now at war, and "in the state of nature…the state of war once begun,

[12] Indeed, you might retaliate even if you think that I was warranted in punishing you but punished you disproportionately. Locke took this possibility seriously: "men being partial to

continues unless…the aggressor offers peace… on such terms as may repair any wrongs he has already done, and secure the innocent for the future" (3:20). But since each sees herself as the innocent party, it is difficult for individuals to come to such terms, and war continues (Nozick 1974: 11-12, Parry 1978: 59; Bruner 2018).

The existence of such feuds—in which punishment begets counter-punishment which begets counter-counter-punishment—occurs in experimental settings closely analogous to the circumstances just mentioned: individuals have an opportunity to engage in multiple rounds of (counter-)punishment, they are aware of who (counter-)punished them last time, and they accept different rules as confirmed by behavior and self-report (Nikiforakis et al. 2012). But one might wonder why such feuds would not be prevented in the state of nature by others engaging in coordinated punishment of their participants, in much the way they can police antisocial counter-punishment under conditions of agreement. Now, sometimes, this might occur. If the overwhelming majority agrees about who has done wrong, they may coordinate punishment and enforce peace. But not all cases will be like this: sometimes there will be significant disagreement over which feudist is in the right. And in these cases, one of two things will happen. Either neither feudist will mobilize anyone to her cause, in which case the feud will go on unperturbed, or both feudists will mobilize supporters, in which case the feud will escalate into war between the groups of supporters. Thus, to the extent that moral disagreements are likely to arise in the state of nature, peace is insecure. Any time an individual or group punishes someone for a perceived moral violation, there is a risk of this escalating into a feud between that individual or group and another. And such feuds, once begun, continue.

---

themselves, passion and revenge is very apt to carry them too far, and with too much heat, in their own cases" (ST 9:125). See fn. 10 above.

Moral disagreements need not be particularly far-reaching to render peace insecure. For even if we suppose that people in the state of nature generally agree about the content and application of morality, so long as there is the occasional case in which individuals disagree, there is a risk that, in such cases, the aforementioned pattern of escalation will occur: individuals or groups line up on different sides of this disagreement, and war ensues. Peace therefore becomes less secure the wider the range of moral disagreement, since this provides more opportunities for conflicts to arise. And if disagreement is broad enough, individuals will no longer be able to predict what others take to be violation of the law of nature, and will lose all trust that their interpretation of the law of nature will be complied with or enforced. At this point, once enforcement is "irregular and uncertain," distrust and war will spread in an all-too-familiar way (*ST* 9:127). Indeed, experiments confirm that, in the face of uncertainty or "noise" about whether individuals have complied with or violated rules—such that if you comply with or violate the rules, there is a non-trivial chance that I think you did the opposite—the ability of punishment to stabilize cooperation rapidly deteriorates (Grechenig et al. 2010; see also Kingsley 2016). And widespread moral disagreement clearly produces such uncertainty.

Thus, under conditions of moral disagreement, the best-case scenario in the state of nature is that disagreement is fairly contained, but peace remains insecure because conflicts may break out over points of moral disagreement. And the worst-case scenario is that disagreement is prevalent, and that so too is war.

6. The Proper Remedy

One notable feature of this account of conflict is that it no longer relies on the presence of dominators. Peace is insecure even among a population of moderates, so long as they

disagree about morality. Indeed, unlike other accounts of Locke's state of nature which trace the source of conflict to deficient moral motivation (Ashcraft 1968: 906; Colman 1983: 185), we can now see that, under conditions of disagreement, ramping up moral motivation may only make matters worse: the higher the cost individuals are willing to pay in order to punish perceived wrongdoing, the more often disagreements over whether such punishment is warranted will lead to war. So we find that moral disagreement drives conflict in the state of nature. Moderates who agree about morality may keep dominators in check through collectively enforcing morality. But moderates who disagree cannot even maintain peace among themselves, since conflict will erupt over such disagreements.

In his political writings, Locke emphasizes bias or partiality to such an extent that it is unsurprising to find many interpretations on which this is the only factor leading to moral disagreement (Colman 1983: 183-185; Simmons 1991: 317; Bruner 2018; though see Parry 1978: 59; Gaus 2018). But Locke himself explicitly acknowledges various other sources of disagreement: "the great variety of Opinions concerning Moral Rules… naturally flows…[from] the different sorts of Happiness [individuals] have a Prospect of" (*EHU* 3:6) as well as from differences in their "Education, Company, and Customs" (3:8). And in recent years, Rawls has enumerated further "burdens of judgment" that lead even impartial individuals to disagree about morality when permitted to judge for themselves (1999: 56-57). Given this variety of factors leading to moral disagreement in the state of nature, and our analysis of how such disagreements produce conflict, we can now understand why Locke insists that "civil government is the proper remedy for the inconveniencies of the state of nature, which must certainly be great, where men may be judges in their own case" (ST 2:13). Moral disagreement is the problem of the state of nature, and government is the solution.

I cannot here tackle the tangled question of how individuals in the state of nature

might succeed in erecting a government (*L* 17; *ST* 8; Nozick 1974: Part I; Hampton 1986: ch. 5; Bruner 2018). Instead, let me end with a brief though somewhat stylized contrast of two ways such a government, once established, might secure peace. The first was suggested by Hobbes, who was also deeply worried about the tendency of disagreements to lead their parties "to blows" in the absence of "some arbitrator or judge to whose sentence they will both stand" (*L* 5:3). Due to this concern, Hobbes insisted that government must be empowered to eliminate all disagreements in society, or—since beliefs "are not voluntary…and consequently fall not under obligation" (40:2)—at least to minimize them and prevent their public expression. Government must therefore have "the whole power of prescribing the rules whereby every man knows what… actions he may do" (18:10), of "hearing and deciding all controversies which may arise concerning law (either civil or natural) or fact" (18:11), and of establishing a state religion to which all must publicly profess (42). It must wield a powerful apparatus of state censorship, deciding "what opinions and doctrines are averse, and what conducing, to peace; and consequently, on what occasions, how far, and what men are to be trusted withal, in speaking to multitudes of people, and who shall examine the doctrines of all books before they be published" (18:9).

Most of us recoil from Hobbes's authoritarian solution to the problem of disagreement, yet our discussion reveals its seemingly impeccable logic. If moral disagreement leads to conflict, and if the "burdens of judgment" imply that agreement "can be maintained only by the oppressive use of state power" (Rawls 1999: 37) then the state must wield such power to maintain peace. Thankfully, Locke recognized a hole this argument, offering a liberal alternative to Hobbes's authoritarianism. He recognized that even if moral disagreement leads to war in the state of nature, a government might solve this problem not by enforcing agreement over all contentious issues, but instead by severing the

connection between disagreement and conflict even while allowing such disagreement to persist and to be publicly expressed.

As our analysis of the state of nature has revealed, it is not moral disagreement per se, or even the expression of such disagreement, that produces conflict. Rather, moral disagreement results in conflict when it leads individuals to enforce divergent understandings of what morality requires, and therefore to feuding and a lack of assurance that any given individual's understanding of morality will be complied with or enforced. Under conditions of moral agreement, these difficulties do not arise because moderates can coordinate their enforcement of a universally recognized system of rules that all interpret and apply in the same way, and may wield their enforcement power to prevent deviant instances of punishment. But since there exist no such system of rules under conditions of disagreement, a government is needed to create it as well as to interpret and apply it. To provide assurance, it must therefore promulgate, publicly adjudicate, and reliably enforce an "*established*, settled, known *law*" (*ST* 9:124). And to prevent feuding, it must obtain a monopoly on punishment by requiring that each individual "wholly gives up" his "power of punishing… as he thought fit" which he enjoyed in the state of nature (9: 130).[13] But this is entirely compatible with

[13] Under conditions of disagreement, government is therefore needed to provide a shared "normative classification" scheme as well as an "authoritative steward" for adjudicating disputes over its interpretation and application (Hadfield and Weingast 2013: 8, 9). But there may be special circumstances of what we may call *near agreement* in which it can reliably enforce such laws without providing a centralized enforcement agency, because the law sufficiently aligns with individuals' moral convictions that individuals willingly coordinate their enforcement of it when directed to by the government. For example, in medieval Iceland, "the only government official… was an individual known as the Law Speaker" who

such laws providing individuals with a wide range of freedom to think, speak, and act as they please, so long as they do not act on their disagreements by violating this law in general or enforcing their own understandings of morality in particular.

Now, in light of people's moral motivation, the efficacy and stability of such a regime may at least require individuals to view it as falling within some acceptable range. Locke himself held that people would rebel against a government that departed too far from their understanding of the law of nature or of who has a right to rule, engaging in a sort of coordinated punishment of "tyrants" and "usurpers" (*ST* 17-19). And a wealth of contemporary evidence—drawn from both experimental settings (Hopfensitz and Reuben 2009; Baldassarri and Grossman 2011) and the real world (Tyler 1990; Barrett and Gaus forthcoming)—confirms that individuals are indeed less motivated to comply with, and more motivated to resist, rules they view as immoral or illegitimate. But it does not follow that a government must enforce agreement even within some acceptable range, or that it must wield a gargantuan power to enforce its laws despite widespread moral motivation pointing in the contrary direction. Instead, Locke hoped that precisely because a liberal state provides space for individuals to live as they choose without fear of others enforcing their divergent conceptions of morality against them, diverse individuals may all be able to agree—or at least not disagree too sharply—that this is indeed the "business of civil government," even while they disagree about issues falling outside of government's purview (Locke 1983: 26).

---

served to recite and interpret rules that were enforced in decentralized fashion" (Hadfield and Weingast 2013: 13). Thanks to an anonymous reviewer for urging me to address such issues.

Though Locke focused on religious tolerance in particular, in recent years, Rawls and other "public reason liberals" have attempted to revise and extend Locke's solution far beyond what Locke himself envisioned, outlining political arrangements that allow those who disagree about a wider range of moral, religious, and philosophical issues to live together on mutually acceptable terms without conflict or oppression (Rawls 1999). This is not the place to go into the details of such views, each of which faces its own difficulties (Vallier 2018). For our purposes, the essential point is that even if Locke is right that moral disagreement is what leads to a need for government, we must nevertheless avoid Hobbes's error of assuming that a government empowered to enforce agreement is the only solution. Instead, the proper remedy may be a government that allows individuals who disagree about morality to live together on peaceful and cooperative terms—terms that these same disagreements preclude them from achieving in the sate of nature.

REFERENCES

Ashcraft, R. 1968. Locke's state of nature: historical fact or moral fiction? *American Political Science Review* 62: 898-915.

Axelrod, R. 2006. *The Evolution of Cooperation.* New York: Basic Books.

Baldassari, D. and G. Grossman. 2011. Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences* 108: 11023-11027.

Barrett, J. and G. Gaus. Forthcoming. Laws, norms, and public justification: the limits of law as an instrument of reform. In *Public Reason and the Courts*, ed. S. A. Langvatn, W. Sadurski, & M. Kumm. Cambridge University Press.

Boyd, R., H. Gintis and S. Bowles. 2010. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* 328: 617-620.

Bruner, J.P. 2018. Locke, Nozick and the state of nature. *Philosophical Studies* Online First: 1-22.

Chaudhuri, A. 2011. Sustaining cooperation in laboratory experiments: a selective survey of the literature. *Experimental Economics* 14: 47-83.

Chung, H. 2015. Hobbes's state of nature: a modern Bayesian game-theoretic analysis. *Journal of the American Philosophical Association* 1: 485-508.

Cinyabuguma, M, T. Page and L Putterman. 2006. Can second-order punishment deter perverse punishment? *Experimental Economics* 9: 265-279.

Colman, J. 1983. *John Locke's Moral Philosophy*. Edinburgh: Edinburgh University Press.

Denant-Boemont, L, D. Masclet and C.N. Noussair. 2017. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory* 33: 145-167.

DeScioli, P. 2016. The side-taking hypothesis for moral judgment. *Current Opinion in Psychology* 7: 23-27.

Dodds, G.G. and D.W. Shoemaker. 2002. Why we *can't* all just get along: human variety and game theory in Hobbes's state of nature. *Southern Journal of Philosophy* 40: 345-374.

Egas, M. and A. Riedl. 2008. The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences* 275: 871-878.

Fehr, E. and S. Gächter. 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90: 980-994.

Fehr, E. and S. Gächter. 2002. Altruistic punishment in humans. *Nature* 415: 137-140.

Fischbacher, U., S. Gächter and E. Fehr. 2001. Are people conditionally cooperative? evidence from a public goods experiment. *Economic Letters* 71: 397-404.

Gaus, G. 2018. Locke's liberal theory of public reason. In *Public Reason in Political Philosophy,* ed. P.N. Turner and G. Gaus, 163-183. New York: Routledge.

Gauthier, D.P. 1986. *The Logic of Leviathan.* New York: Oxford University Press.

Grechenig, K, A. Nicklisch and C Thöni. 2010. Punishment despite reasonable doubt—a public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies* 7: 847-867.

Hadfield, G.K. and B.R. Weingast. 2013. Law without the state: legal attributes and the coordination of decentralized collective punishment. *Journal of Law and Courts* 1: 3-34.

Hampton, J. 1986. *Hobbes and the Social Contract Tradition*. Cambridge: Cambridge University Press.

Hermann, B., C. Thöni and S. Gächter. 2008. Antisocial punishment across societies. *Science* 319: 1362-1367.

Hobbes, T. 1994 [1651]. *Leviathan [L]*, ed. E. Curley. Indianapolis: Hackett.

Hopfensitz A. and E. Reuben. 2009. The importance of emotions for the effectiveness of social punishment. *Economic Journal* 119: 1534-1559.

Kavka, G.S. 1986. *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press.

Kingsley, D.C. 2016. Endowment heterogeneity and peer punishment in a public good experiment: cooperation and normative conflict. *Journal of Behavioral and Experimental Economics* 60: 49-61.

Kogelmann, B. and B.G. Ogden. Enough and as good: a formal model of Lockean first appropriation. *American Journal of Political Science* 62: 682-694.

Ledyard, J.O. 1995. Public goods: some experimental results. In *Handbook of Experimental Economics*, ed. J. Kagel and A. Roth, 111-193. Princeton: Princeton University Press.

Lloyd, S.A. 2009. *Morality in the Philosophy of Thomas Hobbes*. Cambridge: Cambridge University Press.

Locke, J. 1975 [1690]. *An Essay Concerning Human Understanding [EHU],* ed. P.H. Nidditch. New York: Oxford University Press.

Locke, J. 1980 [1690]. *Second Treatise of Government [ST]*, ed. C.B. Macpherson. Indianapolis: Hackett.

Locke, J. 1983 [1689]. *A Letter Concerning Toleration*, ed. J.H. Tully. Indianapolis: Hackett.

Locke, J. 1997 [1692]. *Ethica A*. In *Locke: Political Essays,* ed. M. Goldie, 318-319. Cambridge: Cambridge University Press.

Macpherson, C.B. 1962. *The Political Theory of Possessive Individualism*. New York: Oxford University Press.

Nikiforakis, N. 2008. Punishment and counter-punishment in public good games: can we really govern ourselves? *Journal of Public Economics* 92: 91-112.

Nikiforakis, N, C.N. Noussair and T. Wilkening. 2012. Normative conflicts and feuds: the limits of self-enforcement. *Journal of Public Economics* 96: 797-807.

Ostrom, E., J. Walker and R. Gardner. 1992. Covenants with and without a sword: self-governance is possible. *American Political Science Review* 86: 404-417.

Parry, G. 1978. *John Locke*. London: George Allen & Unwin.

Rawls, J. 1999. *Political Liberalism*. New York: Columbia University Press.

Sheridan, P. 2007. Pirates, kings and reasons to act: moral motivation and the role of sanctions in Locke's moral theory. *Canadian Journal of Philosophy* 37: 35-48.

Simmons, A. J. 1989. Locke's state of nature. *Political Theory* 17: 449-470.

Simmons, A.J. 1991. Locke and the right to punish. *Philosophy & Public Affairs* 20: 311-349.

Tyler, T.R. 1990. *Why People Obey the Law*. New Haven: Yale University Press.

Vallier, K. 2018. Public justification. *Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. https://plato.stanford.edu/entries/justification-public/.

Vanderschraaf, P. 2006. War or peace?: a dynamical analysis of anarchy. *Economics and Philosophy* 22: 243-279.