Janus-Faced Justification*

Gerald Gaus

What a field-day for the heat
A thousand people in the street
Singing songs and carrying signs
Mostly say, hooray for our side
~Steven Stills

1 INTRODUCTION

First and foremost, public reason is a project of reconciliation. It commences from the undeniable fact that reasonable people disagree about moral matters, and seeks a basis for sharing a common view of the demands of morality, or at least justice. To some of us, the importance of such a project in our morally-divided societies is so obvious that we are taken aback by how little interest most political philosophers and, indeed citizens, have in moral reconciliation. While crowds are hurling moral accusations at each other in the street, the philosopher in her study dreams of a social order built around her own vision of justice. The title of Kevin Vallier's recent formulation of reconciliation through public reason poses the rhetorical question whether our moral and political lives must be a war.[1] At one point, I thought not; I now believe that in some respects it is inevitable that sharp conflict and a lack of reconciliation — sometimes leading to outright hostility — are part and parcel of our moral life. This does not undermine public reason's core reconciliation project but, rather, requires that we understand reconciliation and conflict in the broader context of the moral life of a society.

I sketch the outlines of such a broader conception in the following pages. Section 2 reminds us of the necessity of reconciliation: it is not an instinctive moral pacifism, but a recognition that morality has a fundamental role in human social life, and humans are individualistic as well as social creatures. Section 3 connects this role to public justification: public justification is not a tempting option (like the chrome on a '57 Chevy Bel Air), but intrinsic to an expected role of morality in our social life. Section 4 then ties justification to the practice of accountability in the form of what I shall call *justification-as-advocacy*. We then get to (what I see as) a surprising conclusion in section 5: the very

*Some material from section 3 is drawn from a paper written with Jacob Barrett; my thanks to him for permission to use it here, as well as comments on an earlier draft.

[1] Kevin Valier, *Must Politics be War? Restoring our Trust in the Open Society* (Oxford: Oxford University Press, 2018).

nature of the ideal of a shared cooperative morality underlying the public reason project gives rise to the feature of morality that so vexes public reason philosophers: moral justification as inducing conflict rather than reconciliation.

2 THE NECESSITY OF RECONCILIATION

It is sometimes remarked that morality — complete with its judgments and condemnations — cannot be necessary for human cooperation because, after all, highly social and cooperative non-humans animals do not possess morality.[2] However, as is now widely recognized, except for the social insects, the complexity and intensity of human cooperation is of different orders of magnitude than other species; along with the social insects humans are deemed "ultrasocial" species. In many ways the unique feature of humanity is that we have achieved an intensity of social integration approached only by the social insects, but unlike them, human cooperators do not typically share significant genetic relatedness.[3] In marked contrast to the social insects, humans have strong tendencies to self-interest, which in some ways is the oldest, most basic, layer of our psychology.[4] As Hobbes taught us, as opposed to bees, our first and strongest inclination is often to secure our own interests even when this seriously impairs cooperation.[5] Yet despite this we are ultrasocial. The dominant hypothesis today is that the key mechanism for the evolution of our unique type of ultra-sociability has been morality, in particular the ability of humans to internalize shared moral rules to guide both their own behavior and to monitor the behavior of others.[6] When members of a group share moral rules they develop common expectations of what others will do in social contexts, and what others expect them to do; they know when

[2] Peter DeScioli, "The Side-taking Hypothesis for Moral Judgment," *Current Opinion in Psychology*, vol. 7 (2016): 23–27.

[3] See my "The Egalitarian Species," *Social Philosophy and Policy*, vol. 31 (Spring 2015): 1-27. Many "eusocial" insects, such as ants, bees, and wasps are haplodiploid – a female has two alleles but a male only one; insect groups composed largely of such sisters have a degree of genetic relatedness approaching .75, whereas human siblings have a .5 relatedness, and first cousins only .125. After first cousins, human genetic relatedness falls away quickly. The importance of relatedness, of course, is that kin-altruism can explain altruism in the form of one phenotype sacrificing itself for another phenotype who is of a similar genotype. See W.D. Hamilton, "The Genetical Evolution of Social Behaviour: I," *Journal of Theoretical Biology*, vol. 7 (1964): 1-16.

[4] As Peter Richerson and Robert Boyd put it, "we are imperfect and often reluctant, though often very effective cooperators." "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.

[5] Thomas Hobbes, *Leviathan*, edited by Edwin Curley (Indianapolis: Hackett, 1994), pp. 108ff.

[6] There are a host of works defending different versions of this hypothesis. See, e.g., Philip Kitcher, *The Ethical Project* (Cambridge, MA: Harvard University Press, 2011), chap. 2; Christopher Boehm, *Moral Origins* (New York: Basic Books, 2012), chaps. 1-3; Michael Tomasello, *A Natural History of Morality* (Cambridge, MA: Harvard University Press, 2016), pp. 2, 37ff.

others will hold them responsible for violations (more on this anon), and the likely consequences of being held responsible.[7] We thus can count on each other in cooperative contexts.[8] Punishment that is closely bound to rule violation renders sanctions predicable and reliable (as opposed, say to punishment that aims at deterrence).[9] The ability to be guided by normative rules is so important to human social life that many believe that there is a "quasi-instinctive" human capacity to acquire such rules.[10]

There was a time when political and legal philosophers thought that the legal system and its punishments could, pretty much on its own, perform this task of supplying the framework of shared cooperative expectations and behaviors. Recent empirical work, however, has cast strong doubt on this: in lieu of supporting norms, laws are typically ignored or evaded. And when the law conflicts with shared norms it can trigger normative opposition to, and contempt for, law.[11] Increased punishment cannot compensate: sanctions that are perceived as inconsistent with normative expectations or personal normative beliefs are apt to be resisted.[12] Rather than smooth social cooperation, a system that chiefly depends on legal punishment is likely to be characterized by high levels of cheating conjoined with high levels of ineffective punishment. Indeed, in the absence of normative agreement, rather than stabilizing cooperation, punishment can induce feuds.[13]

[7] I explored these functions in *The Order of Public Reason* (Cambridge: Cambridge University Press, 2011), chap. 3. See also Cristina Bicchieri, *The Grammar of Society* (Cambridge: Cambridge University Press, 2006), esp. chap. 1.

[8] P. Kyle Stanford argues that this need to know whom to count on is sufficiently important to drive the "objectification" of morality: those who do not accept our morality are outside our system of accountability. "The Difference Between Ice Cream and Nazis: Moral Externalization and the Evolution of Human Cooperation," *Behavioral and Brain Sciences*, vol. 41 (2018): 1-13. See further section 5 below.

[9] That punishment does not aim at deterrence does not imply that it is not an effective deterrent. See my essay, "Retributive Justice and Social Cooperation" in *Retributivism: Essays on Theory and Practice*, edited by Mark D. White (Oxford: Oxford University Press, 2011): 73-90.

[10] See Hugo Mercier and Dan Sperber *The Enigma of Reason* (Cambridge, MA: Harvard University Press, 2017), p. 71. See also Tomasello, *A Natural History of Morality,* pp. 6, 65ff, 107ff; Joseph Henrich, *The Secret of Our Success* (Princeton: Princeton University Press, 2016), pp. 156, 191, 320ff.

[11] See Jacob Barrett and Gerald Gaus, "Laws, Norms, and Public Justification: The Limits of Law as an Instrument of Reform," available at http://www.gaus.biz/Barrett-Gaus.pdf

[12] See, for example, Samuel Bowles and Herbert Gintis, *A Cooperative Species: Human Reciprocity and its Evolutio*n (Princeton: Princeton University Press, 2011), pp. 26ff; Astrid Hopfensitz and Ernesto Reuben, "The Importance of Emotions for the Effectiveness of Social Punishment," *The Economic Journal*, vol. 119 (2009): 1534–1559.

[13] Nikos Nikiforakis, Charles N. Noussair and Tom Wilkening, "Normative Conflict and Feuds: The Limits of Self-Enforcement, *Journal of Public Economics,* vol. 96 (2012): 797–807.

Ultra-social creatures of our kind thus confront a basic quandary. As philosophers from Hobbes to Rawls have noted, each tends to favor understandings of moral rules that advance her interests,[14] a commonsense belief for which there is strong experimental support.[15] So by its very nature, human social life is characterized by the necessity of a shared morality and a tussle over what these common rules will be. But if this tussle goes too far our social existence impaired. Consequently, from hunter-gatherers onwards, a critical part of social life has been to somehow channel these divergent interests and views into consensual decisions.[16] Thus, we might say, from the very first, human morality has been a reconciliation project.

3 THE NECESSITY OF PUBLIC JUSTIFICATION

To some extent we must reconcile: it is not enough for some to simply declare what the rules will be, they must be widely embraced. Because a critical function of moral rules is to promote the firm shared expectations about behavior so necessary to intense cooperation, in a well-functioning system agents must be *sensitive* to the demands of the rules in the sense that they are willing to forgo personal gains to comply with the shared rules. On Bicchieri's influential analysis of a social norm (or, we might say, moral rule),[17] a person whose personal normative convictions support it will be more sensitive to the its requirements: he will be willing to pay a greater personal cost in order to adhere to it, and so will be more likely to comply with it even in the absence of the threat of sanctions.[18] Bicchieri explains:

> Sensitivity to a norm refers to how much a person adheres to what the norm stands for. Norm sensitivity embodies one's personal reasons for adhering to the norm. A highly sensitive individual could list several good, important, reasons why a particular norm should be enforced, whereas an individual with low sensitivity, who does not care much about what the norm stands for, may only list the fact that, since the norm is widespread, it makes sense for her to obey it (to avoid the sanctions that transgressions incur). Let us call a person's sensitivity to a particular

[14] Hobbes, *Leviathan,* p. 180; Rawls, *A Theory of Justice*, rev edn. (Cambridge, MA: Belknap Press of Harvard University Press, 1999), pp. 171-72, 195-6.

[15] See, e.g. Cristina Bicchieri and Alex Chavez, "Norm Manipulation, Norm Evasion: Experimental Evidence," *Economics and Philosophy,* vol. 29 (July 2013): 175-98; DeScioli, "The Side-taking Hypothesis for Moral Judgment," p. 25.

[16] Herbert Gintis, Carel van Schaik, and Christopher Boehm, "*Zoon Politikon:* The Evolutionary Origins of Human Political Systems," *Current Anthropology,* vol. 56 (June 2015): 327-53.

[17] For the relation between my understanding of a "social moral rule" and Bicchieri's understanding of a social norm, see *The Order of Public Reason,* pp. 163-72. For present purposes, they can be treated as essentially equivalent.

[18] In Bicchieri's formal and empirical work, the sensitivity variable (k) measures a person's tendency to forego monetary gains in order to comply with a fairness norm. Bicchieri, *The Grammar of Society*, pp. 52-54.

> norm, n, k_n. For example, a person who is not very convinced of the advisability of child marriage will have very low sensitivity to that norm (in other words a very low k_n), whereas a person who is convinced that that child marriage is the best way to protect a child's honor will be highly sensitive to the norm.[19]

A person who is sensitive to a social norm is one who believes there are many "good reasons" for adhering to and, presumably, enforcing the norm. A person who is highly sensitive to a norm — who is willing to follow it even at considerable cost to herself — is likely to be one for whom it is justified in the sense that the norm "stands for" or promotes the things she cares about.[20] Let us call this:

> *The Justification Effect:* one's sensitivity (k) to a moral rule/ norm tends to rise as its justification increases, where justification depends on the coherence of the rule/norm with one's own personal normative convictions.

This is only a minimal type of justification: a moral rule is justified to a person if it aligns with her personal normative convictions. Such justification does not interrogate the grounds of those convictions — whether a person's normative convictions are, say, themselves based on badly-grounded beliefs.[21] Still, greater minimal justification tends to induce greater sensitivity.

Bicchieri also points to a more demanding notion of justification.[22] One can come to recognize that one's personal normative convictions are themselves not well-grounded, for example when we realize that our empirical beliefs and normative commitments do not cohere.[23] Bicchieri and Hugo Mercier thus argue:

> Inconsistencies are typically the occasion for belief change. When inconsistent beliefs are detected, the mind tries to determine which can be most easily rejected in order to reduce the inconsistency…. Arguments take a belief that the listener accepts — the premise — and show her that this belief is inconsistent with the rejection of the argument's conclusion. When a good

[19] Bicchieri, *Norms in the Wild* (Oxford: Oxford University Press, 2016), p. 165.

[20] One can view these things a person cares about as "commitments." See, Amartya Sen "Rational Fools" in his *Choice, Welfare and Measurement* (Cambridge, MA.: Harvard University Press, 1982): 84-106. From another perspective, this connects up with what Rawls calls the need for "congruence" between the right and the good. *A Theory of Justice*, pp. 450ff.

[21] The Justification Effect is not uniformly strong throughout the population. Those with greater "reflective autonomy," Bicchieri 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 just because, say, the norm has been in place for a long time, and so will be less sensitive to reasons against it. Bicchieri, *Norms in the Wild,* pp. 166ff.

[22] See Bicchieri and Hugo Mercier, "Self-serving Biases and Public Justifications in Trust Games," *Synthese,* vol. 190 (2013): 909–922; Bicchieri and Mercier, "Norms and Beliefs: How Change Occurs," *Iyyun: The Jerusalem Philosophical Quarterly*, vol. 63 (January 2014): 60–82.

[23] Bicchieri, *Norms in the Wild,* pp. 129-30.

> argument is offered, it is more consistent for the listener to change her mind about the conclusion than to accept the premise while rejecting the conclusion.[24]

While unreflectively one may conclude that one's personal normative beliefs endorse a norm, upon further argumentation or reflection on relevant data, one my come to see either that this is not so, or that one's moral convictions were flawed. When successful this leads to:

> *Robust Public Justification:* a moral rule/norm is robustly justified in a social group *G* if (*i*) at least a large majority of *G* view their personal normative beliefs as giving reasons to hold that everyone in the group ought to act on the moral rule/norm, (*ii*) this conclusion is stable in the light of the amount of reflection on their beliefs, discussion, and exposure to new information that it is reasonable to expect of typical members of *G*.

Obviously clause (*ii*) is contextual, and rather vague.[25] The root idea, though, is that in any given case, a moral rule fails in robust justification if, in the light of the degree of critical reflection and discussion that is appropriate to the group on this matter, they conclude that their personal normative beliefs do not give them reason to endorse it.[26]

As we move from an unjustified moral rule to a minimally justified one, and then on toward robust public justification, not only is the autonomy of the rule follower respected,[27] but the moral rule/norm becomes stronger and more stable in at least three ways. *First*, dissemination of new information is apt to confirm endorsement of the rule: it is not based on insulating error or prejudice from interrogation. *Second*, since personal normative convictions are firmly aligned with the moral rule, individuals are typically more sensitive to it, and informal (and formal) punishment becomes less important as agents become less temped to defect. *Third*, critical reflection and discussion are likely to enhance rather than undermine an individual's normative convictions and therefore

[24] Bicchieri and Mercier, "Norms and Beliefs," p. 69.

[25] For some philosophical cleaning up, see *The Order of Public Reason*, pp. 254-8.

[26] Robust Justification is no mere philosopher's will-o'-the-wisp: it is, essentially, the aim of the Tostan Community Empowerment Program. The program, as conducted in rural Senegal (in villages ranging from 200-500), centers on human rights and democracy education, stressing the exploration of, and deliberation about, the values recognized by the members of the community. Throughout the curriculum, the aim is to examine these ideas in light of the values of the community members. The participants in these classes reflect on human rights and equality (for example, concerning gender norms), often reaching considerable consensus within the group about these values and some of their implications, before going out to engage in further deliberation and discussion with the wider community. See Beniamino Cislaghi, Diane Gillespie and Gerry Mackie, *Values Deliberation and Collective Action in Rural Senegal* (Wallace Global Fund and UNICEF Child Protection Section, 2014); Bicchieri, *Norms in the Wild*, pp. 132ff, 159-69.

[27] In the language of the public reason theorist, each is treated as a free and equal moral person. See *The Order of Public Reason*, chap. 1.

to enhance the individuals' sensitivity to the norm. So robustly justified moral rules are more stable in the face of the spread of information, temptations to defect, and critical reflection and discussion. Their efficacy and stability are not dependent on ignorance or coercion, but on the reflective normative convictions of those they govern.

Thus the importance of some notion of reconciliation in our moral lives. Supposing we start with moral disagreements, if we cannot converge on shared publicly justified rules a critical function of morality is impaired. As I said, reconciliation is not a commitment to moral pacifism, but part and parcel of the basis of human ultrasociality.

4 THE NECESSITY OF ACCOUNTABILITY

Thus far, then, I have argued (*i*) that a critical function of morality is to provide a framework for intense social cooperation through firm shared expectations about cooperative behavior, which sometimes requires one forgoes significant personal gains, but (*ii*) inherent to ambivalent human cooperators is our need to share rules while disagreeing about what they should be; (*iii*) because law and coercion are themselves ineffective at inducing cooperation on common rules, these rules must be internalized by the overwhelming majority; (*iv*) this is accomplished by the public justification of moral rules: when moral rules are robustly publicly justified, agents are sensitive to them, and so shared expectations are strengthened. (*v*) When the group is divided about what rules are normatively endorsed, not only is cooperation impaired, but attempts to enforce some rule via punishment is apt to backfire, producing new conflict.

As I noted at the outset (§2), we tend to interpret moral rules in ways that favor our own interests and, of course, we are often tempted to cheat on them. This can unwind a common morality, and each veers toward acting on self-interest. Thus a critical part of a functioning social morality is a practice of accountability, in which apparent violators are required to justify their actions to others. In important recent experiment Erte Xiao has shown that having to present justifications for one's actions tends to render one more sensitive to the expectations of others.[28] Because this accountability relation tends to enhance sensitivity to the empirical and normative expectations of others, it (*i*) helps keeps people's understandings of the rules coordinated (we are constantly finding out what they expect of us) and (*ii*) checks our inevitable tendency to slide toward self-interested action and interpretations.[29]

The underlying idea of all I have said thus far is that our morality fulfills a necessary role in human social life;[30] we should understand it, and its practice of accountability,

[28] Erte Xiao, "Justification and Conformity," *Journal of Economic Behavior & Organization*, vol. 136 (2017): 15-28.

[29] I have stressed this second point in *The Order of Public Reason*, chap. 4. See also Tomasello, *A Natural History of Morality*, pp. 2, 39ff.

[30] Of course, it is by no means restricted to this role. Automobiles have the role in human life of

in light of that role — as a way for rather independently-motivated agents, with their own aims and interests, to nevertheless enmesh themselves in intense social networks. Now we should understand the place of justification and reasoning-giving in the practice of accountability in a similar, functionality-focused, light, as do Sperber and Mercier:

> By giving reasons to explain and justify yourself, you do several things. You influence the way people read your mind, judge your behavior, and speak of you. You commit yourself by implicitly acknowledging the normative force of the reasons you invoke: you encourage others to expect your future behavior to be guided by similar reasons (and to hold you accountable if it is not). You also indicate that you are likely to evaluate the behavior of others by reasons similar to those you invoke to justify yourself. Finally, you engage in a conversation where others may accept your justifications, question them, and invoke reasons of their own, a conversation that should help you coordinate with them and from which shared norms actually may progressively emerge. Reducing the mechanisms of social coordination to norm abiding, mindreading, or a combination of these two mechanisms misses how much of human interaction aims at justifying oneself, evaluating the reasons of others (either those they give or those we attribute to them), criticizing past or current interactions, and anticipating future ones.[31]

In this matrix of functions, defending oneself and one's reputation with reasons in the face of a charge of noncompliance is critical. As Sperber, Mercier and Haidt all argue, the model of reasoning in these contexts is that of an advocate for one's innocence, not an impartial Kantian inner tribunal.[32] The main role of reasons," according to Mercier and Sperber, "is not to motivate or guide us in reaching conclusions but to explain and justify after the fact the conclusions we have reached."[33] Of course, like any worthwhile advocate, in giving justifications one may see that the jury isn't buying it, and one may have to revise one's plea. As we have seen, when one justifies one becomes sensitive to one's audience and their expectations. But the first attempt is almost always to argue one's case, and to see how many one can get to accept one's plea. Most often one's plea is a not simply a disinterested opinion about what the rules requires — it just so

providing transportation, but this is consistent with stretch limos, Maseratis, Smart cars — and of course '57 Chevy Bel Airs.

[31] See Mercier and Sperber, *The Enigma of Reason*, p. 168.

[32] See Ibid., p. 124; Haidt, *The Righteous Mind*, pp. 81ff; On the Kantian view, see my "Private and Public Conscience" in *Reason, Value, and Respect: Kantian Themes from the Philosophy of Thomas E. Hill, Jr.*, edited by Mark Timmons and Robert Johnson (Oxford: Oxford University Press, 2015): 135-56; Thomas Hill Jr., "Four Conceptions of Conscience" in *NOMOS XL: Integrity and Conscience*, edited by Ian Shapiro and Robert Adams (New York: New York University Press, 1998): 13-52.

[33] Mercier and Sperber, *The Enigma of Reason*, p 112. We need not follow this line of thought to a comprehensive debunking of reasoned argument; in contexts of accusation and justification their analysis of argumentative reasoning is powerful.

happens that this opinion best supports one's wider goals (or, to be more Rawlsian, one's conception of the good). Sperber and Mercier call this a "myside" bias.[34]

5 A STRATEGIC MODEL OF JUSTIFICATION

Once again, let me recap our story. Cooperation requires shared morality; such morality must ground shared expectations that others will behave in a predictable cooperative way; because our interests interreact with our judgments, we tend to use any ambiguity in the rules to favor our own interests and, of course, cheating is an ever-present possibility. We thus need to reconcile, and accept common rules that adequately align with people's normative conviction and interests. Yet, self-interest and disagreement always threaten instability in the shared rules. As a result, it is necessary for us to police each other's performances and hold each other accountable for unjustified violations; and so, a critical human cognitive capacity is to justify oneself to others and show that one should not be held accountable (and possibly punished) for wrongful violations. It is worthwhile stressing how dangerous punishment can be to a person's well-being. In all societies, those deemed to have violated the group's moral rules are, at best, downgraded as possible cooperative partners, and often are subject to various degrees of ostracism.[35] It would be quite remarkable if given all this, in contexts of accountability justification did not take on the role of advocacy for one's position.

To better see some of the possible consequences of justification-as-advocacy, consider a simple model.[36] This very simple model has three key assumptions.

> **(I)** *The Benefits of Expanding Cooperation.* The morality-as-beneficial-cooperative framework view is supposed: every individual sees increasing her network of moralized cooperation as always, to at least some extent, a good thing — and conversely, being excluded from a network is always, to some extent, a bad (of disvalue). Expanding cooperation is by no means all a moral agent is interested in, but it is always an interest. This is the classic evolutionary view, and we are trying to better understand where it might lead us.

[34] Ibid., pp. 218ff.

[35] For a now classic analysis of hunter-gatherer societies and levels of punishment, see Christopher Boehm, *Hierarchy in the Forest* (Cambridge, MA: Harvard University Press, 1999), chaps. 3 and 4. For a famous ethnographic case, see Colin M. Turnbull, *The Forest People* (New York Simon and Schuster, 1962), chap. 5. More generally on the costs of punishment for moral violations, see Stanford, "The Difference between Ice Cream and Nazis."

[36] It is important to stress that models do not seek to describe all possible, or even all important cases, but to enlighten us about some dynamic. They tell a story and, and, as we know, there is always another one to be told. See Ariel Rubinstein, *Economic Fables* (Cambridge: Open Book Publishers, 2012); James Johnson, "Review Essay – Formal Models in Political Science: Conceptual, Not Empirical," *Journal of Politics,* vol. 81(2018) < http://dx.doi.org/10.1086/700590>.

(II) *Moral Disagreement.* We suppose that people differ on what would be the morally best cooperative rule. Because we suppose that moral views and self-interests are intertwined, we suppose that a person would overall benefit more from universal action on her favored rule than the alternative. A person's decision as to what rule is best, we assume, depends on her view of what is morally correct; but since moral judgment is informed with self-interest, and people's justifications track their interests, we can expect people to justify different rules.

(III) *Strategic Advocacy.* This leads us to our third assumption, Peter DeScioli and Robert Kurzban's strategic view of moral disagreement: "Individuals stand to gain by proposing and defending moral rules that benefit themselves"[37] and thus will argue strategically in moral contexts. Consistent with this and Mercier-Sperber's view (§4), it is supposed that a person's response to a charge of non-compliance is to be explained by what, all things considered, most benefits her (with this including moving society to her most favored moral position).

Consider, then, a simple world composed of group *G* with two possible moral rules, and everyone acts on one or the other. Suppose that Alf has performed action φ. Betty, a follower of moral rule R, claims that φ is wrong on her interpretation of R, and she holds him responsible for a violation of R. In our simple model Alf has only two options. He can CONCEDE GUILT, accepting that Betty's is the proper interpretation of R, a correct moral rule. Alternatively, he can DEFEND, rejecting the rebuke, advancing a justification claiming either (*i*) that Betty's is not the proper interpretation of R or (*ii*) that R is not the correct moral rule. Let us combine *i* and *ii* by saying that, when he DEFENDS, Alf claims that φ was not wrong according to moral rule R* which is either the correct interpretation of R or the correct moral alternative to R.

If Alf DEFENDS, he will be excluded from the R cooperative network (call this R_n—the number *n* of *G* individuals who uphold R.). He will have sent a signal to the upholders of R that he dissents, and so that he cannot be relied upon to act on the expectations generated by R, and refuses to be held accountable for its violation. Given Assumption **I**, the larger the R_n network, the higher the costs of non-participation, which we denote as $C(R_n)$. On the other hand, by DEFENDING R* Alf signals his loyalty to the R* network (R^*_{G-n}), the number of people in group *G* who follow rule R* (rather than the *n* who follow R). Denote the benefits of participating in the R* network $B(R^*_{G-n})$. Of course size of network is not decisive: if Alf favors R* a cost of CONCEDING would be remaining in the R network, which offers less favorable terms of moral cooperation (though, presumably, if R is sufficiently larger that can compensate). If he DEFENDS and so joins the R* network, the moral superiority of R* would be an additional benefit (which could

[37] Peter DeScioli and Robert Kurzban, "A Solution to the Mysteries of Morality," *Psychological Bulletin*, vol. 139 (2012): 477–496, at p. 488.

compensate for it being a smaller network). His decision will thus be driven by the overall relative benefits of R and R* to him, and the relative size of the cooperative moral networks that provide these benefits. Alf, then, will choose:

$$\text{DEFEND if:} \quad B(\text{R}^*_{G\text{-}n}) > C(\text{R}_n)$$
$$\text{CONCEDE if:} \quad C(\text{R}_n) > B(\text{R}^*_{G\text{-}n})$$

Case 1: Moral Coordination. Consider three paradigmatic cases. In the first, Alf regularly concludes that $C(\text{R}_n)$ is *much greater* than $B(\text{R}^*_{G\text{-}n})$, as does just about everyone else. Moral diversity (Assumption **II**) still holds (Alf may favor R*), but by far most accept R, and so the size of the R network is so overwhelmingly large that no one wishes to be excluded. Here the members of group *G* overwhelming support the R rule, and so treat any violation as impermissible and warranting moral condemnation. In this case, even if Alf finds R objectionable as a basis for social relations, the sheer number who support it will most likely lead him to CONCEDE, and not seek to justify his φ-ing. As P. Kyle Stanford argues, in such situations morality allows us to easily coordinate our actions: there might be many possible ways of coordinating (R or R*), but when a large majority of our group believes that R is universally correct, it will not tolerate departures from R: to reject R is to endanger one's status as a competent group member.[38] In *Case 1* accountability and justification function as aids to moral coordination. DEFEND would constitute a public signal that one rejects what the group deems to be the demands of moral cooperation. There are few benefits to arguing for a revised rule in this accountability context, and great costs to explicitly breaking with what is understood as morally right. Insofar as Alf uses justification-as-advocacy it will be in the way of CONCEDING, say, to point excusing factors.

Case 2: Choosing Sides. In the second case, the group is divided into, say, three rough types: advocates of R, advocates of R* and a significant "quasi-indifferent" group for whom the values of $B(\text{R}^*_{G\text{-}n})$ and $C(\text{R}_n)$ are either equal or very close to equal. Here, if Alf decides to DEFEND the morality of action φ, members of group *G* choose sides.[39] If

[38] Stanford, "The Difference between Ice Cream and Nazis."

[39] This discussion draws on Peter DeScioli and Robert Kurzban's theory of morality as choosing sides. The basic case in their account is one in which society has a non-moral conflict, and morality is a way for people to choose sides in an impartial way (thus, for example, a brother, appealing to morality, may side against his own kin). Thus, in this way morality helps resolve disputes in a way that is not driven by prestige or kin-relations. Yet they also see that morality as choosing sides leads to fights over the moral rules — we take sides over how disputes are to be settled. The focus here is on this latter case. See DeScioli, "The Side-taking Hypothesis for Moral Judgment;" DeScioli and Kurzban," A Solution to the Mysteries of Morality;" DeScioli and Kurzban, "Morality Is for Choosing Sides" in the *Atlas of Moral Psychology*, edited by Kurt Gray and Jesse Graham (New York: Guilford Publications, 2017): 177-85.

Alf DEFENDS, he will have two aims: (*i*) to signal to R* devotees that he is a reliable member of their network and (*ii*) to convince members of the quasi-indifferent subgroup that impartial considerations are on the side of R*, thus expanding the size of the R* network (and so his own $B(R^*_{G\text{-}n})$). Many in the quasi-indifferent subgroup are apt to be less sensitive (§3) to either R or R*; in their eyes neither manifestly better aligns with their normative convictions or, perhaps, their personal normative convictions render them ambivalent about the rules. Here, should Alf DEFEND, his use of justification-as-advocacy is likely to be *reformist*, arguing to others, especially to the wavering R followers in the quasi-indifferent subgroup, of the superiority of R* and the flaws of R. Convincing them to also adopt R* would, *pro tanto*, increase $B(R^*_{G\text{-}n})$ to him. The existence of the less sensitive quasi-indifferent subgroup is critical for this case: they are most apt to be moved by arguments for or against R and R*. In crafting his justification, Alf will be sensitive to their concerns. As John Tooby and Leda Cosmides argue, if R* is to "climb the ladder of increasingly wide support," the case for it must extend beyond its parochial appeal to a highly committed R* subgroup. This incentivizes presenting impartial and general cases for R*, appealing to wider array views.[40] Thus we see that the critical features of moral argument, impartiality and universalization, are elements of strategic justification (assumption **III**).[41] In *Case 2* moral reconciliation on a publicly justified shared moral rule is entirely possible despite differences in moral ideals. Indeed — and this is the interesting point — such a society tends to produce justifications with wide appeal. We thus arrive at an important hypothesis: societies that have successfully coped with moral diversity at one level may well be those that can continue expanding their moral networks because they have achieved wider-based, more impartial, justifications.[42] In "climbing the ladder" of wider appeal in a diverse society, they have crafted their rules to accommodate greater diversity.

In *Case 2* convergence on a shared rule can also be modelled in terms of dynamic individual choices, even without a change in normative belief or common acceptance of the same arguments: a significant quasi-indifferent subgroup can still help push society on a path to ultimate convergence on one or the other rule.[43] The key here is that $B(R^*_{G\text{-}n})$ increases as the number of people endorsing R* increases while $C(R_n)$ decreases;

[40] John Tooby and Leda Cosmides, "Groups in Mind: The Coalitional Roots of War and Morality," in *Human Morality and Sociality: Evolutionary and Comparative Perspectives*, edited by H. Høgh-Olesen (New York: Palgrave Macmillan, 2010): 191-234 at p. 220.

[41] On the importance of impartiality in getting people to choose your side, see the references to DeScioli and Robert Kurzban's model in the note 39.

[42] See further, *Tyranny of the Ideal*, pp. 173ff.

[43] For details see my "Self-organizing Moral Systems," *Politics, Philosophy and Economics*, vol. 17 (May 2018): 119-147.

even without changing one's evaluation of the inherent morality of R and R*, as some people switch to accepting R*, others' estimate of the relative benefits and costs of CONCEDING or DEFENDING change. To see the dynamic clearly, assume that Charlie, in our quasi-indifferent group is strictly indifferent: the benefits and costs of being in either network are equal. This means that, given (*i*) his relative evaluations of the extent to which the two rules align with his normative commitments, and (*ii*) the numbers now following each rule, the benefits and costs of being in either network are the same. Suppose that originally, perhaps by the slip of a coin, Charlie opted for R, but now that he has witnessed Alf DEFENDING he comes to believe he has misestimated the relative size of the networks, and R* is slightly bigger than he has previously thought. He now believes that R* has grown (so R has shrunk), by Assumption **I** he must believe that R* has become more attractive.[44] That implies that it now is the case that for Charlie the net benefits of being in the R* network exceed the R network: he thus switches to the R* network and henceforth will DEFEND if accused of R violations. Now consider Doris, for whom previously the net benefits of R network were ever-so-slightly greater than the R* network. Charlie's defection to R* can cause her to reevaluate her favored network, also switching to R*, and from there — if there is enough diversity in people's relative evaluations of the two rules[45] — a cascade can take place to R*, becoming the new rule for the group. There is no reason why both this and the Tooby – Cosmides dynamic cannot operate in tandem; as an impartial, broader, justification leads some in our quasi-indifferent group to switch their moral position to rule R*, others may cascade to it on the basis of the second dynamic.[46]

Note here that the very justificatory competency that is critical to a stable shared moral rule (*Case 1*), can be also employed to undermine the current rule and move to a

[44] Some have pointed out that Assumption **I** treats all people as equally important cooperators, and this is an idealization that may well not be the case. First, remember note 36. Anyway, No and Yes. *No*, because Assumption **I** allows highly nonlinear functions relating increases in the number following a rule to how much better that makes the rule. Charlie might, say, place little marginal importance on R* increasing from 1 to 10 in its networks ("Ten times nothing is basically nothing!"), but see much more benefit in a marginal increase from 100 to 110; in this sense the 1-10*th* and the 100-110*th* members are not treated as equally important. *Yes*, because the model treats individuals as interchangeable: the agents are anonymous insofar as their value as cooperators does not depend on special facts about them. The model can be modified to accommodate different values of cooperators by having each person weight the importance of others according to, say, prestige; highly prestigious persons could count for 3, middle for 2, low for 1. Once the weightings are added the model then can be run as described in the text; this sentence in the text remains true so long as everyone counts for something.

[45] For the importance of this diversity for cascades, see my "Self-organizing Moral Systems."

[46] Indeed, they also suggest a cascade dynamic. "Groups in Mind," pp. 224-25. The first basis for change is what public reason theorists typically call a "shared reason" for adopting R* while the second illustrates "convergence" reasoning. It is a mistake to see these as incompatible; we see here how they can interact.

new publicly justified rule. (Obvious recent examples of this sort of change are rules concerning racial and sexual discrimination.) Justification must be able to perform this destabilizing role if a cooperative moral system is to learn and adapt. As recent analyses such as Haidt's, Stanford's and DeScioli-Kurzban's have recognized, any adequate account of morality must be able to induce change as well as provide stability.[47] Moral diversity and conflict may be an engine of moral reform, pointing us toward a new cooperative equilibrium.[48] On the other hand, we should expect continued conflict on many matters. "As moral projects climb the ladder to broader audiences, (being recast and potentially applied to increasingly broad sets of individuals), any given individual will be bombarded with increasing numbers of candidate moral rules."[49] Thus the paradoxical feature of our times: the better our morality is at accommodating diversity the more challenges diversity poses to it.

Case 3: Radical Polarization. Suppose now that *G* is split into only two subgroups, where for one $C(\mathrm{R}_n)$ is much greater than $B(\mathrm{R}^*_{G\text{-}n})$ and for the other $B(\mathrm{R}^*_{G\text{-}n})$ is much greater than $C(\mathrm{R}_n)$. There is no quasi-indifferent group. It is important to stress the two subgroups need not be even approximately equally sized; intensity of valuing may compensate in a person's decision-making for lack of numbers.[50] *Case 3* is likely to occur when members of the two subgroups see the other rule as entirely unacceptable (as I have put it in various places, "ineligible.") The first group will always CONCEDE when faced with a violation of R, the second will always DEFEND (and, *mutatis mutandis*, the opposite with R* violations). Here each subgroup strategically uses justification-as-advocacy to solidify the group's devotion to the rule, not to generate converts from the other: no one is trying to broaden their justifications as a way of appealing to the (non-existent) quasi-indifferent group (cf. *Case 2*). Each subgroup's justifications reinforce its own rules, while viewing the other as morally beyond the pale. Myside bias will be deeply rooted and reinforced. Each preaches to the choir, and each increasingly firmly holds a moral view that simply cannot be embraced by the other group.

If this division is overlapping such that a number of disputes all have the same subgroup boundaries (say, one subgroup strongly advocates moral rules defending gay rights, vegetarianism, environmental-focused moral rules, a woman's right to abort and

[47] See DeScioli, "The Side-taking Hypothesis for Moral Judgment;" DeScioli and Kurzban," A Solution to the Mysteries of Morality;" DeScioli and Kurzban, "Morality Is for Choosing Sides;" Sanford, "The Difference between Ice Cream and Nazis;" Haidt, *The Righteous Mind*, chap. 12.

[48] See further my *Tyranny of the Ideal* (Princeton: Princeton University Press, 2016), pp. 230-40.

[49] Tooby and Cosmides, "Groups in Mind," p. 224.

[50] See Linda J. Skitka, Christopher W. Bauman, and Edward G. Sargis, "Moral Conviction: Another Contributor to Attitude Strength or Something More?" *Journal of Personality and Social Psychology*, vol. 88 (2005): 895–917.

economic redistribution) while the other subgroup strongly opposes these rules, the empirical evidence highlights three possible upshots. The *first*, most dire (and not terribly unlikely), possibility is that inter-group hatred arises, not only hindering cooperation but inducing conflict and, at the extreme, violence.[51] We have to remember that when morality is evoked, the other is not simply different: she is turning her back on MORALITY. Not only condemnations as "immoral" but as "evil" may arise. Moreover, as we have seen, any attempt to punish the other subgroup is likely to invoke counter-punishment.

A second possibility is that the subgroups gravitate apart: given that a perception of another as immoral is associated with less cooperative relations with them,[52] this perhaps should be expected: many such disputes are resolved by self-segregation.[53] As I have argued elsewhere, this may be a social rather than a spatial separation; the subgroups can carve out different social worlds in which, to some extent, their distinctive moralities hold sway.[54] Interestingly, people's views of the morality of an out-group appear to become increasingly "relativistic" as cooperation with them decreases and they operate in different cultural milieus.[55] In these cases people become less apt to insist on accountability relations with others with whom they recognize sharp disagreements and, so, condemnation of their acts becomes less likely.[56]

Lastly, the dispute may eventually be solved by *demoralization*. As people's empirical expectations about the behavior of others is repeatedly disappointed, the existence of *any* rule may be undermined.[57] To be sure this is unlikely to occur with the basic moral rules of cooperation such as those specifying liberty, harm and property. Cooperation without such moral rules is, at best, difficult and imperfect.[58] But on other matters what was formerly a moral requirement may become a personal choice. Allen Buchanan and Russell Powell see moral progress in many "demoralizations" — "Examples include profit-seeking, lending money at interest, masturbation, premarital sex, same-

[51] Robert Böhm, Isabel Thielmann, and Benjamin E. Hilbig, "The Brighter the Light, the Deeper the Shadow: Morality Also Fuels Aggression, Conflict, and Violence," *Behavioral and Brain Sciences*, vol. 41 (2018): 15-16.

[52] Skitka, Bauman, and Sargis, "Moral Conviction."

[53] Stanford, "The Difference between Ice Cream and Nazis," p. 9.

[54] These might be associations or friendship networks. I have argued that rights can bound such protected social spaces. See *The Order of Public Reason*, chap. 7.

[55] See Hagop Sarkissian, John Park, David Tien, Jennifer, Cole Wright and Joshua Knobe, "Folk Moral Relativism," *Mind & Language*, vol. 26 (September 2011): 482–505.

[56] In consociational states characterized by strict cleavages, while elites interact to manage the system, the constituent groups experience decreased reduced relations with each other. Arend Lijphart, *Democracy in Plural Societies* (New Haven: Yale University Press,1977), pp. 48ff.

[57] See Bicchieri, *Norms in the Wild*, chap. 3.

[58] DeScioli and Kurzban," A Solution to the Mysteries of Morality," p. 488.

sex sexual relations, interracial marriage, and (some instances of) civil disobedience."[59] Whether or not these quintessential liberal liberties are matters of progress, they certainly are developments that have allowed those with sharply different perspectives to cooperate without recrimination, though perhaps still with distaste.[60]

6 JANUS-FACED JUSTIFICATION

When moral rules are publicly justified, each internalizes them and expects others to do so as well. Here justification enhances social cooperation and we are at moral peace, as in *Case 1*. But a publicly justified morality includes a practice of accountability, which in turn gives rise to argumentative justification, constructing cases for beneficial moral outcomes. This argumentative justification often leads to moral conflict — or, even as Tooby and Cosmides describe it, "moral warfare."[61] Each has the argumentative resources to press for changes in moral rules: the very competencies required for the practice of a stable morality thus induce instability, giving effective voice to divergent views and interests. Such instability is by no means to be always regretted. Somewhat paradoxically, it can be the driver of moral reform, inducing a more impartial morality, which is more accommodative of wider diversity. Here justificatory conflicts generated by diversity lead to a public morality less fragile in the face of further diversity, expanding the sphere of moral cooperation (*Case 2*). This, I have argued elsewhere, is a great achievement of the open society.[62] On the other hand, justificatory disagreement can crystalize into hostility and indeed, mutual hatred (*Case 3*). Should this crystallization occur — as, alas, it may be doing in parts of our own society — I would hope that political philosophers resist imitating those officers in the First World War who, obsessed by dreams of a glorious victory in the trenches, egged their troops on in a righteous struggle that shredded a cooperative order.

Political Economy & Moral Science/Philosophy
University of Arizona

[59] Allen Buchanan and Russell Powell, *The Evolution of Moral Progress: A Biocultural Theory* (New York: Oxford University Press, 2018), p. 56 and generally chapter 8. Cf. *The Order of Public Reason*, pp. 315-19.

[60] This is probably why "WEIRD" (Western Educated Industrialized Rich Democratic) morality is focused on harm and liberty: it is a solution to the problem cooperative moral order under conditions of moral disagreement. See Haidt, *The Righteous Mind,* Part II.

[61] Tooby and Cosmides, "Groups in Mind," pp. 224-25.

[62] *The Tyranny of the Ideal*, chap. 4.