Normative Conflict and Normative Change

Graham Alexander Noblit,[a,b,*], Gillian K. Hadfield[a,b,c]

[a] *Schwartz Reisman Institute, 661 University Ave Suite 710, M5G 1M1, Toronto, Ontario, Canada*
[b] *Vector Institute, 101 College St Suite 230-4, M5G 1L7, Toronto, Ontario, Canada*
[c] *Faculty of Law, University of Toronto, M5S 3E6, Toronto, Ontario, Canada*

Abstract

Human life is heavily structured by norms, many of which require individuals to adopt net-costly behaviors. Such costly behaviors often generate the non-rivalrous and non-excludable benefits that permit human ecological adaptation. However, models studying the mechanisms that stabilize costly-norms emphasize that their stability need not consider what group benefits the norm confers. Additionally, in a dynamic world, innovations, environmental change, and information revelation imply that what norms are beneficial for a group will change over time. Both of these phenomena emphasize the importance of understanding mechanisms of normative change. Explanations of intra-societal normative change are relatively rare, buy when investigated they have generally failed to identify conditions under which a mutant and more-beneficial behavior can invade a population characterized by a previously stabilized norm and have pointed to group-level selective mechanisms to explain the presence of adaptive norms. We construct a cultural evolutionary game theoretic model to study normative change by means of intra-societal costly-punishment and conflict resolution and identify social differentiation in punishment capacity as a key condition permitting normative change. Importantly, we find that this normative change will be group beneficial, however, the assumptions permitting said result appear to be highly restrictive.

1. Introduction

In contrast to other vertebrates, human life is uniquely structured by norms (Bicchieri, 2006; Boyd, 2019; Roughley and Bayertz, 2019; Searle, 1995). Although debate exists as to how humans evolved to be normative and what a

*Corresponding author
Email addresses: graham.noblit@utoronto.ca (Graham Alexander Noblit,), g.hadfield@utoronto.ca (Gillian K. Hadfield)

proper definition for norms is, scientists across diverse fields including evolutionary anthropology and cultural evolution (Boyd, 2019; Henrich and Ensminger, 2014) and economics (Bicchieri, 2006; Nunn, 2020) emphasize the role of norms and institutions, or packages of norms (Ensminger and Henrich, 2014; Henrich and Ensminger, 2014; Henrich and Muthukrishna, 2021), for determining human psychological, behavioral, and group-level outcomes. A subset of norms are particularly relevant for human adaptation: Costly normative behaviors are behaviors whereby individuals shoulder private-costs exceeding their derived private-benefits and are generally thought to be stabilized by the actions of third-parties Chudek and Henrich (2011). Punishment mechanisms, such as the withdrawal of mutual aid (Panchanathan and Boyd, 2004); sanctions, either costly (Boyd and Richerson, 1992; Boyd et al., 2010) or extractive and self-interested (Bhui et al., 2019); and costly signaling (Gintis et al., 2001) re-align individuals' incentives such that group-members have reason to adopt net-costly behaviors. Importantly, these mechanisms stabilize behaviors regardless of their group-level benefits, or the benefits a norm confers to a focal individual when all group-members adopt it. As such, the above mechanisms are expected to not only stabilize behaviors that are beneficial to individuals but ones that are harmful as well.

Though the above models demonstrate that we should expect groups to be saddled with non-beneficial norms, they also demonstrate that normative change is itself something of a theoretical quandary. Imagine that a population, for whatever reason, has come to be characterized by some normative practice that is demonstrably harmful to all members. Any individual who seeks to "shirk" the norm and adopt a less-costly behavior will be seen as a "defector" by their group co-inhabitants, and quickly punished by all group-members. That we might generally consider only a few mutants to exist simultaneously, this

means that behavioral and normative mutants will be wildly outnumbered in large groups and thus likely to quickly realign their behavior with the group's preferred behaviors. This is true regardless of whether the behavior adopted by the mutants would confer greater benefits to all other group-members than the status quo behavior can. Panchanathan and Boyd (2004) show that in a population where a costly-norm has been stabilized by strategies that withhold (beneficial) mutual aid from norm-breakers, a form of punishment, an alternative costly-norm cannot invade even when said norm provides greater benefits to all group-members, i.e. everyone in the group would in fact prefer that an alternative normative equilibrium be stabilized. The above issue is compounded by the fact that realistic environments are intrinsically dynamic and behaviors that were once adaptive may become maladaptive relative to alternatives should environmental shifts occur (Boyd and Richerson, 1985; Rogers, 1988) or novel innovations emerge.

Although these various issues suggest we should observe only a random distribution of normative benefits, human groups are generally characterized by norms and institutions that are highly adaptive and beneficial (Boyd, 2019; Kelly, 1995), such as stable contributions to public and common goods, sometimes at great personal risk (Mathew and Boyd, 2011). More so, such cooperative behaviors likely extend deep into human cultural and genetic evolution (Boyd and Richerson, 2022), suggesting some mechanism is and has been at play to ensure that beneficial norms are stabilized. The adaptive nature of norms and institutions might reflect individuals' reasoning capacities (Richerson and Henrich, 2012), however, individuals often cannot (or, in some cases, it is impossible for them to be able to) explain the benefits of their practices (Boyd et al., 2013; Henrich, 2002, 2016). Similarly, efforts to consciously implement and improve on institutions and norms often fail (Richerson and Henrich, 2012)

and individuals in experiments often do not predict institutional equilibria correctly (Gurerk et al., 2006), suggesting that human reasoning cannot explain the presence of beneficial and adaptive norms and institutions.

So how do we explain the plethora of beneficial norms and institutions that exist? How did they come to be adapted to highly variable environments? Cultural evolutionists have generally relied on cultural group selection to explain how the observed distribution of norms becomes more adaptive over time. Cultural group selection occurs when behaviors propagate relative to the benefits they confer at the group-level, leading adaptive and beneficial norms to "outcompete" harmful ones (Boyd and Richerson, 2009, 2002, 2009; Henrich, 2004; Richerson et al., 2016). Most theorized group selection mechanisms act at the level of groups and produce changes in an aggregate distribution of equilibria, such as when different norms and institutions lead groups to go extinct at variable rates or heterogeneously succeed in inter-group combat.

Cultural group selection theorists have not extensively explored intra-group normative change for obvious reasons (though see Boyd and Richerson, 2002, for what might be considered an important exception) and have generally assumed that punishment mechanism, which stabilize norms, should they be applied to normative change, will only serve self-interest (Richerson et al., 2016). However, empirical work highlights the historical importance of the political, legal, and generally collective processes organized around normative conflict in determining how societies establish and change their norms (Boehm et al., 1996; Cohen and Middleton, 1967; Hadfield and Weingast, 2013). Such normative conflict can be a source of both group-beneficial (Keyssar, 2009) and extractive (benefiting a subset of group-members at cost to another) normative change (Bello, 2019). Normative conflict deserves increased attention because it has played a key role in social evolution during the Holocene of the past $\approx 12,000$ years

(Flannery and Marcus, 2012; North et al., 2009; Price et al., 1995) and has established important "prosocial" norms and institutions, that broaden access to benefits, characterizing the Leveller and Digger movements in 17^{th} century England (Rees, 2017; Wood, 2017), which partially sought to effectively liberalize and democratize the social and political institutions of the era, efforts by the Roman plebs seeking release from overwhelming debt obligations to the patricians and stronger legal rights (Forsythe, 2005), and collective labor action by Egyptian scribes in the 12^{th} century BCE (Edgerton, 1951). Normative conflict also features in an important socio-behavioral shift in human evolution (Boehm et al., 1993; Boehm, 1999), whereby chimpanzee dominance hierarchies, which likely serve the private interests of a single dominant male or small coalition of males, came to be upended by a coalition of the weak that coooperatively enforces relatively "egalitarian" norms and institutions in order to "level" differences across individuals and preventing bullying and extractive behaviors by the strong. In these cases, factions of individuals sought to stabilize alternative and mutually exclusive normative equilibria, rules that all societal members would adhere to; more so, these rules sought changes that would impose on their groups what many (Richerson et al., 2016) might consider now to be adaptive and sometimes "prosocial" institutions. Political and legal ethnography similarly identifies cases where agents differ in their normative vision for their communities, in their "opinions of good citizenship and impropriety" (Panchanathan and Boyd, 2004, p.501) and seek to coerce group-members' cooperation with and adherence to said vision, coercing and resisting normative change (Bailey, 2001; Barth, 1959; Boehm et al., 1993; Boehm, 1999; Ensminger and Knight, 1997; Flannery and Marcus, 2012; Hayden, 2011, 2014, 2016; Kirch, 2010; Price et al., 1995, 2010; Wiessner, 2002).

We argue that intra-societal processes producing and resolving normative

conflict necessitate further direct study because they may offer pathways to group-beneficial institutions absent cultural group selection mechanisms. To study the role of normative change in human social evolution, we construct a stylized model that permits agents engage in normative conflict and for conflict to be settled in a way that permits normative change. Following Boyd and Richerson (1992) we permit agents to impose punishments-costs on one another at cost to themselves. We study a decentralized form of sanctioning (Boyd and Richerson, 1992) because it provides an analytically tractable way to examine the role of social-differentiation on the evolution of norms while simultaneously capturing the capacity for individuals to confer costs on one another and for such conflict to potentially be resolved. Costly-punishment (Boyd and Richerson, 1992) likely does not describe how all costly-norms are stabilized. Various mechanisms can stabilize costly-behaviors in n-player settings (Bhui et al., 2019; Boyd et al., 2010; Chudek and Henrich, 2011; Gintis et al., 2001; Henrich et al., 2015; Panchanathan and Boyd, 2004) and empirical evidence suggests a multiplicity of costly-norm enforcement mechanisms exist (Henrich et al., 2006; Herrmann et al., 2008; Marlowe et al., 2008; Marlowe, 2009; Mathew and Boyd, 2011; Smith et al., 2018). Our model provides a baseline for comparison with future and alternative models and is not meant to provide an overarching theory of normative change.

We model conflict resolution akin to as modeled by Boyd and Richerson's (1992) where we assume no political or legal institutions exist to resolve normative conflict. Instead, we ask when resident strategies' best-response, once punished by a socially-differentiated mutant, is to abandon their current normative behavior and adopt the mutant's preferred behavior, permitting normative change. We show that permitting normative conflict and resolution means that the stability criterion of a norm now includes the group-level benefit terms

associated with the behaviors in question. While this result appears to offer an evolutionary ratchet to explain the presence of adaptive norms and institutions absent any group-selection forces, it stems from two important and highly restrictive assumptions. First, it requires a principle we refer to as "social differentiation" whereby conflict with a normative mutant, which we refer to as a "Change Agent", be costlier to members of the population as compared to conflict with the entire rest of their group. Second, it requires that this social differentiation occur *as a function* of what public good behaviors the agents adopt, i.e. which potential norms they seek to stabilize, and that this social differentiation effect occur privately, i.e. not reflect any coordinated forms of punishment.

2. Model

Our model follows standards in evolutionary game theory. Assume groups of size n are sampled from an effectively infinitely sized population. After each round, groups continue to exist with probability ω. Within each round, two stage games exist. First, there is a linear public goods game (Boyd and Richerson, 1988, 1992) whereby agents can contribute to public good 1 or 2. These public goods entail different benefits to group-members. Each individual's marginal contribution to either public good $i \in 1, 2$ confers a benefit $\frac{b_i}{n}$ to all group-members. This means if s individuals contribute to 1 and t individuals contribute to 2, each individual will receive $b_1 \frac{s}{n} + b_2 \frac{t}{n}$. Each public good is also characterized by a distinct private cost of contributing, c_i where i again denotes which public good an agent adopts.

The second stage game is a punishment game where agents can adopt a cost to confer punishment on other agents. Importantly, to study the role of more realistic punishment technologies on the evolution of norms, we assume that the

cost of punishment and of being punished depend on an interaction between what public good an agent adopts as well as who the social partner they are engaging with it. An agent will shoulder a cost, k_i^j, to impose a cost of p_j^i on another agent where $(i, j) \in \{(1, 2), (2, 1)\}$ and denotes the public good behavior that a punisher and recipient of punishment have adopted. For instance, an agent that adopts public good 2 and who punishes an agent contributing to public good 1 will confer a punishment p_2^1 at cost k_1^2 to themselves. Technically, our superscript notation is superfluous because we never examine punishment across norms (i.e. we never study a 2 adopter that punishes other 2 adopters). The above means that if a strategy initially adopts behavior 1 and punishes behavior 2 contributors in the first round before adopting behavior 2 and punishing behavior 1 contributors in the second round, its punishment parameters will have changed. Effectively, the conflict parameters are something like "private goods" in that they depend on the agent's actions alone. Correspondingly they cannot represent any coordinated (Boyd et al., 2010; Hadfield and Weingast, 2012) (or generally strategic) form of punishment that would require multiple group members to take specific actions.

Agents are defined by inherited strategies defining actions in each stage game (decision contingency). A strategy's replication depends on its fitness, which is proportionate to its payoffs. We assume that implementation error rates are identical between strategies and that their role is the evolution of punishment is well understood (Boyd and Richerson, 1992).

3. Results

3.1. Permanent Conflict

We initially study an ecology defined by two strategies:

- **Stubborn 1** ($\mathbf{S_1}$): Play behavior 1, punish any agent that does not adopt behavior 1
- **Stubborn 2** ($\mathbf{S_2}$): Play behavior 2, punish any agent that does not adopt behavior 2

These strategies represent the coexistence of two strategies attempting to impose their preferred behaviors on their group. In such an ecology, any group consisting of a mix of these strategies will engage in ongoing and permanent conflict because (1) each strategy punishes any strategy that fails to adhere to its normative ideal (which behavior it adopts) and (2) there is no mechanism for conflict resolution. Correspondingly, no focal agent will ever see its group-members switch which public good they contribute to, however, fitness differences among strategies may lead to the proliferation of one public good relative to another on cultural evolutionary timescales.

The invasion criterion of $\mathbf{S_2}$ when $\mathbf{S_1}$ is common (alternatively, $\mathbf{S_1}$'s stability criterion) is given by (proof in Appendix A):

Proposition 1. *When strategies cannot resolve normative conflict then a new norm can replace a resident norm so long as any change in private net-costs that the normative mutant experiences in adopting its preferred behavior exceeds the cost of conflict it faces when punishing and being punished by the $n-1$ resident norm adherents. For a 2-bearing strategy invading a population of 1-bearing strategies, the invasion criterion takes the following form:*

$$\overbrace{(c_1 - \frac{b_1}{n}) - (c_2 - \frac{b_2}{n})}^{\Delta C_{1\to 2}\text{: Change in net-costs that 2-bearing Change Agent experiences for adopting behavior 2}}$$

$$>$$

$$\underbrace{(n-1)\,(p_{1\to 2} + k_{2\to 1})}_{\text{Cost of conflict that Change Agent experiences due to violating norm 1}}$$

For a 1*-bearing strategy invading a population of* 2*-bearing strategies, the invasion criterion takes the following form:*

$$\overbrace{(c_2 - \frac{b_2}{n}) - (c_1 - \frac{b_1}{n})}^{\Delta C_{2\to1}\text{:Change in net-costs that 1-bearing Change Agent experiences for adopting behavior 1}}$$

$$>$$

$$\underbrace{(n-1)\,(p_{2\to1} + k_{1\to2})}_{\text{Cost of conflict that Change Agent experiences due to violating norm 2}}$$

A resident norm can only be replaced by an alternative norm if any change in private net-costs of adopting the novel behavior exceeds the cost of conflict that the agent adopting the alternative norm experiences for this normative violation. This criterion is essentially identical to Boyd and Richerson's (1992) seminal result that punishment can stabilize behaviors arbitrary of the benefits they confer (no b term is present). 1 only differs in that (1) it considers a change in costly behaviors as opposed to merely free-riding on one and (2) the change in net-costs of cooperating, which we label ΔC, must now exceed not only the cost of being punished by $n-1$ residential norm-bearing agents (p) as well as the cost the mutant adopts when she punishes said $n-1$ agents (k).

We might expect the above criterion to generally be unlikely to be met and thus the process it formalizes could rarely be expected to explain normative change. It is unrealistic, particularly when we consider ethnographically described cases of group-based material punishment, which include tolerated theft and destruction of food stores, ostracism (and the withdrawal of mutual aid), executions, beatings, and monetary fines (Amadiume, 2015; Boehm, 1999; Briggs, 1971; Henrich and Henrich, 2014; Humphrey, 1942; Mathew and Boyd, 2011, 2014; Smith, 1961), to assume that the change in net-private costs from adopting a novel behavior could offset the cost of conflict with the resident strategies.

Additionally, this model critically assumes that no strategy will ever respond to punishment by adopting an alternative norm, i.e. no agent is permitted to acquiesce to punishment and instead, conflict continues indefinitely and is never resolved. The presence of permanent conflict suggests that evolution may select for strategies which optimize who they engage in conflict with, if conflict with different strategies entails distinct costs. This argument is akin to pointing to the fact that punishing strategies cannot directly invade defecting populations that fail to respond to punishment with cooperation; such recalcitrant defectors are in fact stable to invasion by punishing cooperators. In order to invade such defecting populations, punishing strategies instead raise the fitness of defecting strategies that *do* respond to punishment relative to recalcitrant defectors and then invade the "contrite" defecting population (Boyd and Richerson, 1992). In a sense, punishing strategies can be thought of as "niche constructors" (Odling-Smee et al., 2003) that seek to construct social environments conducive to their own proliferation. A similar process may permit a normative mutant to coerce its own evolution.

3.2. Conflict Resolution by Punishment

We now consider what evolutionary incentives the presence of costly normative conflict produces when what public good behaviors a strategy adopts impact their conflict parameters (the p and k terms). We assume the existence of a population where behavior 1 was historically stabilized by punishment, i.e. is the only present behavior, and we ignore the mixed equilibrium for punishing strategies located in Boyd and Richerson (1992); this is a conservative assumption because such a mixed equilibrium would be more permissive of normative change. Assuming that behavior-2 adopting mutants repeatedly arise (we label any agent that adopts a minority public good behavior as "Change Agents"; Change Agent 2 denotes a strategy that breaks the most frequent norm in the

population of adopting behavior 1 by adopting behavior 2 instead), perhaps due to behavioral errors and cognitive constraints (Boyd and Richerson, 1992), we ask when a strategy that responds to $\mathbf{S_2}$ punishment by adopting behavior 2 has a higher fitness than the resident strategy, $\mathbf{S_1}$, which ignores $\mathbf{S_2}$ punishment and continues to play behavior 1 (proof in Appendix B):

Proposition 2. *Strategies that respond to Change Agent* 2 *punishment by adopting behavior* 2 *(Fair-Weather Friend:* $\boldsymbol{S_{1\to 2}}$*) will replace the resident norm* 1 *playing strategy (*$\boldsymbol{S_1}$*, which ignores Change Agent punishment, if conflict with the Change Agent when failing to adopt behavior* 2 *sufficiently exceeds the cost of conflict with* $n-2$ *residential group members for adopting behavior* 2*:*

$$\overbrace{(c_1 - \frac{b_1}{n}) - (c_2 - \frac{b_2}{n})}^{\Delta C_{1\to 2}:\text{Change in private net-costs that Fair-Weather Friend experiences for adopting behavior 2}} + \underbrace{(p_2^1 + k_1^2)}_{\text{Cost of conflict with Change Agent}}$$

$$>$$

$$\underbrace{(n-2)(p_1^2 + k_2^1)}_{\text{Fair-Weather Friend's experience of conflict with } n-2 \text{ behavior 1 group-members after adopting 2}}$$

When the above criterion is met, a 1 playing strategy's best response is to adopt behavior 2 in order to avoid $\mathbf{S_2}$ punishment. This criterion will be met so long as the cost of conflict with the Change Agent ($\mathbf{S_2}$; $p_2^1 + k_1^2$), which is what a norm 1 playing strategy experiences when it refuses to adopt behavior 2, is costlier than conflict with the $n-2$ resident norm 1 playing strategies, $(n-2)(p_1^2 + k_2^1)$, which is what a Fair-Weather Friend strategy that ends up adopting behavior 2 experiences.

Proposition 2 makes a critical point: How a behavior interacts with an agent's punishment parameters, both in terms of the costs it can confer to other agents as well as the costs other agents bear when engaging in conflict with it, will impact the evolution of norms. This interaction effect must be "private"

and occur at the level of the individual; any change in punishment parameters as a result of adopting the novel behavior must result from an individual mutant's actions alone and not represent any more sophisticated coordinated forms of punishment (Boyd et al., 2010; Hadfield and Weingast, 2012). However, correspondingly, the above means that it is only for the subset of behaviors wherein adopting them alters agents' punishment capacities in a way that satisfies the criterion in Proposition 2. This means that relatively strong restrictions exist in terms of what normative change can be supported in order for conflict to induce normative change on the timescale of a group's existence. Alternatively, our model implies that historical normative conflict will select for those norms that confer to individuals punishment parameters defined in Proposition 2.

Now that we have identified the conditions under which normative change is possible on the timescale of a group's existence, we identify the the stability criterion of a previously stabilized norm, behavior 1, which is also the invasion criterion of behavior 2. In assuming that Proposition 2 holds, the ecology of strategies no longer includes **Stubborn 1** but rather **Fair-weather Friend**:

1. **Fair-weather Friend ($\mathbf{S_{1\rightarrow 2}}$)**: Play behavior 1, punish any agent that does not adopt behavior 1. If punished by a behavior 2 playing agent, adopt behavior 2 for all future rounds and punish any behavior 1 playing agent.
2. : **Stubborn 2 ($\mathbf{S_2}$)**: Always play behavior 2, punish any agent that does not adopt behavior 2.

Because we assume Proposition 2 holds, now when **Stubborn 2** encounters the resident norm 1 adherents, a single round of normative conflict will occur before the norm 1 adherents adopt behavior 2. When can a new norm, behavior 2 invade and begin to replace a pre-existing norm, behavior 1 (proof in Appendix C):

Proposition 3. *A Change Agent that can coerce cooperation with its preferred norm (Proposition 2 holds) will begin to replace resident norm adherents so long as the lifetime benefits of normative change plus any differences in the net cost of cooperation when adopting the new behavior exceed the initial cost of normative conflict:*

$$\overbrace{(c_1 - \frac{b_1}{n}) - (c_2 - \frac{b_2}{n})}^{\Delta C_{1\to 2}\ \textit{Change in net costs when switching from 1 to 2}} + \underbrace{\frac{\omega}{1-\omega}\left[(b_2 - c_2) - (b_1 - c_1)\right]}_{\Delta B_{1\to 2}\textit{: Change in induced benefits from switching from behavior 1 to 2}}$$

$$>$$

$$\underbrace{(n-1)(p_1^2 + k_2^1)}_{\textit{Cost of conflict to 2-bearing mutant}}$$

If this criterion is not met then norm 1 *is stable to invasion.*

The above criterion portrays normative conflict as akin to an "investment". So long as the cost of that investment to the Change Agent mutant, $(n-1)(p_1^2 + k_2^1)$, is less than the "return", or the induced life-time *change in benefits* that the Change Agent experiences from coercing its $n-1$ group members to shirk 1 and adopt 2, $\Delta B_{1\to 2} = \frac{\omega}{1-\omega}\left[(b_2 - c_2) - (b_1 - c_1)\right]$, then norm 2 will invade norm 1. Otherwise, norm 1 is stable to invasion. In effect, we show that the stability of a norm, behavior 1, which has been stabilized by punishment (i.e. is fixed in a population), depends on the relative group level net-benefits of two behaviors.

Does the stability of norm 2, when it is common, similarly depend on a change in net-benefits (proof in Appendix C)?

Proposition 4. *When norm* 2 *is dominant in the population, and norm* 2 *bearing agents can coerce cooperation with that norm (Proposition 2 holds), behavior* 1 *will begin to replace behavior* 2 *so long as the punishment that the behavior*

1 *mutant experiences for adopting behavior* 1 *is exceeded by any change in net-costs that it experiences when switching to behavior* 1*:*

$$\overbrace{(n-1)(p_2^1+k_1^2)}^{\textit{Cost of conflict to behavior 1 mutant}} > \underbrace{(c_2-\frac{b_2}{n})-(c_1-\frac{b_1}{n})}_{\textit{Change in net-costs to mutant in switching from behaviors 2 to 1}}$$

If this criterion is not met then norm 2 *is stable to norm* 1 *invasion.*

This criterion states that a resident behavior is stable so long as the cost to a mutant of engaging in political conflict with it's entire group, $(n-1)(p_2^1+k_1^2)$, exceeds the change in private net-costs to from switching to behavior 1 from the resident norm, $(c_2-\frac{b_2}{n})-(c_1-\frac{b_1}{n})$, then a behavior is stable. This criterion closely matches the stability criterion of Boyd and Richerson's (1992) and states that the stability of a behavior when stabilized my costly-punishment in no sense depends on the benefit of that behavior. In effect, punishment allows the evolution of cooperation (or anything else) in group settings.

Finally, there is the possibility of a stable interior equilibrium (proof in Appendix C).

Proposition 5. *Should the criteria in both Propositions 3 and 4 fail to be met, i.e. neither the benefits of normative change exceed the cost of conflict to* 2*-bearing agents nor does the cost of conflict to a* 1*-bearing mutant exceed the change in private net-costs the mutant experiences in adopting behavior* 1*, then a stable mixed internal equilibrium consisting of a small proportion of* $\boldsymbol{S_2}$ *exists.*

4. Discussion

We have demonstrated that, in contrast to previous results Boyd and Richerson (1992), the stability of a norm at times depends on what group-level benefit it affords (Proposition 3). Somewhat strangely, we have shown that this is not

true for another norm (Proposition 4). How do we explain this distinction in our results?

The key is to understanding the apparent conflicting results is to direct attention to the role of punishment and the nature of conflict resolution in the model. Because no institutions exist to permit normative conflict to be settled, it must be settled by means of punishment. Proposition 2 permits conflict to be resolves and provides the criterion under which behavior 1 adherents are incentivized to adopt behavior 2 once punished by behavior 2 adherents. However, when this criterion is true, which depends on conflict with 2-bearing agents to be far costlier than conflict with a group of 1-bearing agents, behavior 2 adherents are never incentivized to adopt behavior 1 once punished. A focal strategy in other words will only, within the scale of its lifetime, switch from behavior 1 to behavior 2, but never vice versa from 2 to 1. Correspondingly, only 2-bearing individuals experience the benefits of their group-members adopting behavior 2 whereas 1-bearing individuals cannot ever coerce cooperation with behavior 1 and never see the benefits of their group-members adopting behavior 1. Should behavior 2 be established in a population, selection may permit 1 to invade when the criterion in Proposition 4 will be true, but because no 1-bearing mutant ever coerces her group to switch from behavior 2 to behavior 1, the group benefits of behavior 1 are irrelevant.

A clearer statement of our key result (Proposition 3) is that so long as (1) adopted behaviors interact privately to determine strategy's punishment capacities and (2) a given norm grants a sufficiently high increase in punishment capacity compared to a resident norm, as described by the criterion in Proposition 2, *then* the stability of a resident norm depends on the life-time change in group benefits associated with adopting the alternative norm. So long as the life-time change in net-group benefits exceeds the cost of engaging in conflict to

the Change Agent mutant coercing cooperation with a novel norm, then the resident norm will be replaced. Proposition 4 then and the stability criterion of the "alternative" norm, behavior 2, does not include any group level benefit terms because no mutant agent is ever able to coerce cooperation with their preferred norm, behavior 1. If neither a Change Agent mutant nor the resident strategies can coerce cooperation, than we are simply left in the world of ongoing conflict (Proposition 1) whereby again, normative change only arises from evolutionary dynamics and the stability of a norm does not depend on group level benefits. The stability criterion result above indicates that understanding how normative conflict comes to be resolved and the game-theoretic properties may have far reaching consequences for what sorts of norms emerge. Correspondingly, future study should be directed towards conflict resolution institutions, drawing on political and legal anthropology and history.

What are the historical implications of these results? We feel there are two obvious insights to our results and a critical weakness. First, that when conflict is resolved by means of punishment, normative conflict will only select for behaviors inducing increased group level benefits from among the set of behaviors that effectively determine strategies' punishment capacities. If the potential norms do not impact punishment capacity to make the criterion in Proposition 2 true, then group-benefits will be irrelevant to the resident norm's stability criterion and no evolutionary ratchet exists to explain the emergence of adaptive norms and institutions. By and large, punishment technologies have been only abstractly formalized in the cultural evolutionary literature. Our model suggests that that is a mistake and that important details about how punishment occurs and relates to bodies of knowledge and material technologies may have deep repercussions for normative evolution. This claim is a subset of our previously mentioned point that future research should explore variation

in how normative conflict is resolved beyond the punishment mechanism we explore, because those characteristics will similarly impact normative change. We have modeled only one specific manner in which normative conflict can be resolved, punishment, and one specific form of punishment at that, where strategies' punishment parameters stem from what public good behavior they privately adopt. This captures a situation where, for instance, in a patriarchal society, norms that restrict women's access and involvement in public places also prevent women from garnering the requisite skills, resources, experience, and psychological traits that would permit them to impose large punishment costs on others. However, other methods of normative conflict resolution exist, and should they resolve conflict stochastically, then Proposition 4 would still entail group benefits, permitting a residential norm to be replaced by a less-adaptive one.

Second, note that the criteria in Propositions 2 and 3 are relatively difficult to meet. Ethnographic evidence suggests that when groups of individuals confer punishment, such as when punishment takes the form of theft and destruction of food stores, ostracism (and the withdrawal of mutual aid), executions, beatings, and monetary fines (Amadiume, 2015; Boehm, 1999; Briggs, 1971; Henrich and Henrich, 2014; Humphrey, 1942; Mathew and Boyd, 2011, 2014; Smith, 1961), the costs of normative conflict can be rather high. Correspondingly, one should suspect that relatively few behaviors will pass the initial "sieve" defined by the criterion in Proposition 2 whereby the costs of conflict with an individual normative mutant exceeds the cost of conflict with an entire group of resident-norm adherents, and the requirement that punishment parameters reflect the private adoption of a single behavior. However, should punishment take coordinated forms and punishment parameters be nonlinear in the number of punishers or the relative size of punishing coalitions, Propositions 2 and 3 would likely be

easier to satisfy. This again emphasizes the importance of understanding specific punishment technologies and how the adoption of behaviors impacts the ability of agents to confer and resist punishment costs.

In light of the importance of how social and material technologies impact the ability of strategies to both confer and resist punishment costs, hierarchical behaviors that construct social distance within a population will likely impact normative evolution. Correspondingly, the critical drawback of our model is that it assumes that the alternative behaviors we study identically impact all agents' payoffs within the population. For most cases of interest, such an assumption is likely far too strong. A more realistic model would permit the population (or meta-population) to consist of agents whose cultural fitness are heterogeneously impacted by the two behaviors we study. Such heterogeneity is part and parcel of evolution, arising from the divergent "preferences" and interests that accompany standard genetic and cultural evolutionary processes. Divergent evolutionary interests in sexual reproduction and sexual selection is an ongoing topic of study in evolutionary biology (Fisher et al., 2013; Parker, 1978, 2006) and critical to understanding hominin evolution (Hrdy, 1997, 1999; Mulder and Rauch, 2009). The culturally variable sexual division of labor denotes specialization into distinct economic roles and likely produces heterogeneous preferences for goal-seeking agents (Bird, 1999; Bird and Codding, 2015; Hadfield, 1999; Waguespack, 2005). Among foragers, such variation may lead individuals to prefer different collective movement decisions, which are of obvious relevance to group-living organisms generally (Krause and Ruxton, 2002), group-compositions, and distributions of individuals' time-allocations. Similarly, individuals are characterized by a relatively large number of distinct biological and psychological traits that may be either culturally or genetically inherited or arise through the stochastic processes defining individual develop-

ment and learning. That such inter-individual variation exists and produces normative conflict is already central to a major theory of human evolution: How the hominin lineage transitioned from chimpanzee dominance hierarchies, in which a single male or small-coalition of males enforce their private reproductive interests by imposing behaviors on group-members, to human foragers, who often violently enforce norms and practices that limit competition, hierarchy, and exploitation by self-aggrandizing and bullying individuals seeking to elevate their private interests above others' (Boehm et al., 1993; Boehm, 1999). Should multiple evolving subpopulations exist within a meta-population and be characterized by distinct cultural fitness orderings, then it is unlikely that normative conflict will permit the evolution of group-beneficial norms (Richerson et al., 2016). In these cases, such as would represent when men and women prefer distinct normative equilibria and individuals largely learn their behaviors from within their gender or sex, the norms that evolve by means of conflict will likely exclusively benefit a given type (gender or sex) within the meta-population. Such dynamics would likely heavily influence the dynamics of group-selection as well, potentially limiting its capacity to select for adaptive norms.

References

C. Bicchieri, The Grammar of Society, Cambridge University Press, New York, NY, 2006.

R. Boyd, A Different Kind of Animal: How Culture Transformed Our Species, Princeton University Press, Princeton, 2019.

N. Roughley, K. Bayertz (Eds.), The Normative Animal?: On the Anthropological Significance of Social, Moral, and Linguistic Norms, Oxford University Press, New York, NY, 2019.

J. R. Searle, The Construction of Social Reality, The Free Press, New York, 1995.

J. Henrich, J. Ensminger, Theoretical foundations: The coevolution of social norms, intrinsic motivation, markets, and the institutions of complex societies, in: J. Ensminger, J. Henrich (Eds.), Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective, Russell Sage Foundation, New York, NY, 2014, pp. 19–44.

N. Nunn, The historical roots of economic development, Science 367 (2020) eaaz9986. doi:10.1126/science.aaz9986.

J. Ensminger, J. P. Henrich (Eds.), Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective, Russell Sage Foundation, New York, New York, 2014.

J. Henrich, M. Muthukrishna, The origins and psychology of human cooperation, Annual Review of Psychology 72 (2021) 207–240. doi:10.1146/annurev-psych-081920-042106.

M. Chudek, J. Henrich, Culture-gene coevolution, norm-psychology and the emergence of human prosociality, Trends in Cognitive Sciences 15 (2011) 218–226. doi:10.1016/j.tics.2011.03.003.

K. Panchanathan, R. Boyd, Indirect reciprocity can stabilize cooperation without the second-order free rider problem, Nature 432 (2004) 499–502. doi:10.1038/nature02978.

R. Boyd, P. J. Richerson, Punishment allows the evolution of cooperation (or anything else) in sizable groups, Ethology and Sociobiology 13 (1992) 171–195. doi:10.1016/0162-3095(92)90032-Y.

R. Boyd, H. Gintis, S. Bowles, Coordinated punishment of defectors sustains cooperation and can proliferate when rare, Science 328 (2010) 617–620. doi:10.1126/science.1183665.

R. Bhui, M. Chudek, J. Henrich, How exploitation launched human cooperation, Behavioral Ecology and Sociobiology 73 (2019) 78. doi:10.1007/s00265-019-2667-y.

H. Gintis, E. A. Smith, S. Bowles, Costly signaling and cooperation, Journal of Theoretical Biology 213 (2001) 103–119. doi:10.1006/jtbi.2001.2406.

R. Boyd, P. J. Richerson, Culture and the Evolutionary Process, University of Chicago Press, Chicago, 1985.

A. R. Rogers, Does biology constrain culture, American Anthropologist 90 (1988) 819–831. doi:10.1525/aa.1988.90.4.02a00030.

R. L. Kelly, The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways, Smithsonian Institution Press, Washington, DC, 1995.

S. Mathew, R. Boyd, Punishment sustains large-scale cooperation in prestate warfare, Proceedings of the National Academy

of Sciences 108 (2011) 11375–11380. doi:10.1073/pnas.1105604108. arXiv:https://www.pnas.org/content/108/28/11375.full.pdf.

R. Boyd, P. J. Richerson, Large-scale cooperation in small-scale foraging societies, Evolutionary Anthropology: Issues, News, and Reviews 31 (2022) 175–198. doi:10.1002/evan.21944.

P. Richerson, J. Henrich, Tribal social instincts and the cultural evolution of institutions to solve collective action problems, Cliodynamics 3 (2012). doi:10.21237/C7clio3112453.

R. Boyd, P. J. Richerson, J. Henrich, The cultural evolution of technology: Facts and theories, in: P. J. Richerson, M. H. Christiansen (Eds.), Cultural Evolution: Society, Technology, Language, and Religion, The MIT Press, 2013, p. 0. doi:10.7551/mitpress/9780262019750.003.0007.

J. Henrich, Decision-making, Cultural Transmission and Adaptation in Economic Anthropology, in: J. Ensminger (Ed.), Theory in Economic Anthropology, AltaMira Press, Walnut Creek, CA, 2002, pp. 251–295.

J. Henrich, The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter, Princeton University Press, Princeton, New Jersey, 2016.

O. Gurerk, B. Irlenbusch, B. Rockenbach, The competitive advantage of sanctioning institutions, Science 312 (2006) 108–111. doi:10.1126/science.1123633.

R. Boyd, P. J. Richerson, Culture and the evolution of human cooperation, Philosophical Transactions of the Royal Society B: Biological Sciences 364 (2009) 3281–3288. doi:10.1098/rstb.2009.0134.

R. Boyd, P. J. Richerson, Group beneficial norms can spread rapidly in a structured population, Journal of Theoretical Biology 215 (2002) 287–296. doi:10.1006/jtbi.2001.2515.

R. Boyd, P. J. Richerson, Voting with your feet: Payoff biased migration and the evolution of group beneficial behavior, Journal of Theoretical Biology 257 (2009) 331–339. doi:10.1016/j.jtbi.2008.12.007.

J. Henrich, Cultural group selection, coevolutionary processes and large-scale cooperation, Journal of Economic Behavior and Organization 53 (2004) 3–35. doi:10.1016/S0167-2681(03)00094-5.

P. Richerson, R. Baldini, A. V. Bell, K. Demps, K. Frost, V. Hillis, S. Mathew, E. K. Newton, N. Naar, L. Newson, C. Ross, P. E. Smaldino, T. M. Waring, M. Zefferman, Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence, Behavioral and Brain Sciences 39 (2016) 1–71. doi:10.1017/S0140525X1400106X.

C. Boehm, C. Antweiler, I. Eibl-Eibesfeldt, S. Kent, B. M. Knauft, S. Mithen, P. J. Richerson, D. S. Wilson, Emergency decisions, cultural-selection mechanics, and group selection [and comments and reply], Current Anthropology 37 (1996) 763–793. doi:10.1086/204561.

R. Cohen, J. Middleton, Comparative Political Systems: Studies in the Politics of Pre-Industrial Societies, American Museum Sourcebooks in Anthropology, Published for the American Museum of Natural History [by] the Natural History Press, Garden City, N.Y., 1967.

G. K. Hadfield, B. R. Weingast, Law without the state: Legal attributes and the coordination of decentralized collective punishment, Journal of Law and Courts 1 (2013) 3–34. doi:10.1086/668604.

A. Keyssar, The Right to Vote: The Contested History of Democracy in the United States, Basic Books, New York, NY, 2009.

W. Bello, Counterrevolution: The Global Rise of the Far Right, Practical Action Publishing, Warwickshire, 2019.

K. Flannery, J. Marcus, The Creation of Inequality: How Our Prehistoric Ancestors Set the Stage for Monarchy, Slavery, and Empire, Harvard University Press, Cambridge, 2012.

D. C. North, J. J. Wallis, B. R. Weingast, Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History, Cambridge University Press, Cambridge, 2009. doi:10.1017/CBO9780511575839.

T. D. Price, G. M. Feinman, G. M. Feinman, T. D. Price (Eds.), Foundations of Social Inequality, Fundamental Issues in Archaeology, Springer US, Boston, MA, 1995. doi:10.1007/978-1-4899-1289-3.

J. Rees, The Leveller Revolution: Radical Political Organisation in England, 1640-1650, Verso Books, 2017.

A. Wood, Riot, Rebellion and Popular Politics in Early Modern England, Bloomsbury Publishing, 2017.

G. Forsythe, A Critical History of Early Rome: From Prehistory to the First Punic War, University of California Press, 2005.

W. F. Edgerton, The strikes in Ramses III's twenty-ninth year, Journal of Near Eastern Studies 10 (1951) 137–145. arXiv:542285.

C. Boehm, H. B. Barclay, R. K. Dentan, M.-C. Dupre, J. D. Hill, S. Kent, B. M. Knauft, K. F. Otterbein, S. Rayner, Egalitarian behavior and reverse dominance hierarchy [and comments and reply], Current Anthropology 34 (1993) 227–254. doi:10.1086/204166.

C. Boehm, Hierarchy in the Forest: The Evolution of Egalitarian Behavior, Harvard University Press, Cambridge, 1999.

F. G. Bailey, Stratagems and Spoils: A Social Anthropology of Politics, Westview Press, Boulder, Colorado, 2001.

F. Barth, Political Leadership among Swat Pathans, University of London, Athlone Press, London, 1959.

J. Ensminger, J. Knight, Changing social norms: Common property, bridewealth, and clan exogamy, Current Anthropology (1997). doi:10.1086/204579.

B. Hayden, Big man, big heart? The political role of aggrandizers in egalitarian and transegalitarian societies, in: D. R. Forsyth, C. L. Hoyt (Eds.), For the Greater Good of All: Perspectives on Individualism, Society, and Leadership, Palgrave Macmillan US, New York, 2011, pp. 101–118. doi:10.1057/9780230116269_7.

B. Hayden, The Power of Feasts: From Prehistory to the Present, Cambridge University Press, New York, NY, 2014.

B. Hayden, Feasting in Southeast Asia, University of Hawaii Press, Honolulu, Hawaii, 2016.

P. V. Kirch, How Chiefs Became Kings: Divine Kingship and the Rise of Archaic States in Ancient Hawai'i, University of California Press, 2010.

T. D. Price, G. M. Feinman, G. M. Feinman, T. D. Price (Eds.), Pathways to Power, Fundamental Issues in Archaeology, Springer New York, New York, NY, 2010. doi:10.1007/978-1-4419-6300-0.

P. Wiessner, The vines of complexity: Egalitarian structures and the institutionalization of inequality among the Enga, Current Anthropology 43 (2002) 233–269. doi:10.1086/338301.

J. Henrich, M. Chudek, R. Boyd, The big man mechanism: How prestige fosters cooperation and creates prosocial leaders, Philosophical Transactions of the Royal Society B: Biological Sciences 370 (2015) 20150013. doi:10.1098/rstb.2015.0013.

J. Henrich, R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. Tracer, J. Ziker, Costly punishment across human societies, Science 312 (2006) 1767–1770. doi:10.1126/science.1127333.

B. Herrmann, C. Thoni, S. Gachter, Antisocial punishment across societies, Science (New York, N.Y.) 319 (2008) 1362–1367. doi:10.1126/science.1153808. arXiv:1101.2204.

F. W. Marlowe, J. C. Berbesque, A. Barr, C. Barrett, A. Bolyanatz, J. C. Cardenas, J. Ensminger, M. Gurven, E. Gwako, J. Henrich, N. Henrich, C. Lesorogol, R. McElreath, D. Tracer, More 'altruistic' punishment in larger societies, Proceedings of the Royal Society B: Biological Sciences 275 (2008) 587–592. doi:10.1098/rspb.2007.1517.

F. W. Marlowe, Hadza cooperation: Second-party punishment, yes; third-party punishment, no, Human Nature 20 (2009) 417–430. doi:10.1007/s12110-009-9072-6.

K. M. Smith, T. Larroucau, I. A. Mabulla, C. L. Apicella, Hunter-gatherers maintain assortativity in cooperation despite high levels of residential change and mixing, Current Biology 28 (2018) 3152–3157.e4. doi:10.1016/j.cub.2018.07.064.

R. Boyd, P. J. Richerson, The evolution of reciprocity in sizable groups, Journal of Theoretical Biology 132 (1988) 337–356. doi:10.1016/S0022-5193(88)80219-4.

G. K. Hadfield, B. R. Weingast, What is law? A coordination model of the characteristics of legal order, Journal of Legal Analysis 4 (2012) 471–514. doi:10.1093/jla/las008.

I. Amadiume, Male Daughters, Female Husbands: Gender and Sex in an African Society, Zed Books, London, 2015.

J. Briggs, Never in Anger: Portrait of an Eskimo Family, Harvard University Press, Cambridge, MA, 1971.

J. Henrich, N. Henrich, Fairness without punishment: Behavioral experiments in the Yasawa Island, Fiji, in: J. Ensminger, J. Henrich (Eds.), Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective, Russel Sage Foundation, New York, 2014, pp. 225–258.

N. D. Humphrey, Police and tribal welfare in Plains Indian cultures, Journal of Criminal Law and Criminology 33 (1942) 147–161. doi:10.2307/1136421.

S. Mathew, R. Boyd, The cost of cowardice: Punitive sentiments towards free riders in Turkana raids, Evolution and Human Behavior 35 (2014) 58–64. doi:10.1016/j.evolhumbehav.2013.10.001.

R. J. Smith, The Japanese rural community: Norms, sanctions, and ostracism, American Anthropologist 63 (1961) 522–533. doi:10.1525/aa.1961.63.3.02a00050.

F. J. Odling-Smee, K. N. Laland, M. W. Feldman, M. W. Feldman, Niche Construction: The Neglected Process in Evolution, Princeton University Press, Princeton, 2003.

M. Fisher, J. R. Garcia, R. S. Chang (Eds.), Evolution's Empress: Darwinian Perspectives on the Nature of Women, Oxford University Press, Oxford, 2013.

GA. Parker, Sexual selection and sexual conflict, in: M. S. Blum, N. A. Blum (Eds.), Sexual Selection and Reproductive Competition in Insects, Academic Press, London, UK, 1978, pp. 123–166.

GA. Parker, Sexual conflict over mating and fertilization: An overview, Philosophical Transactions of the Royal Society B: Biological Sciences 361 (2006) 235–259. doi:10.1098/rstb.2005.1785.

S. B. Hrdy, Raising Darwin's consciousness, Human Nature 8 (1997) 1–49. doi:10.1007/s12110-997-1003-9.

S. B. Hrdy, Mother Nature: A History of Mothers, Infants, and Natural Selection, Pantheon Books, New York, NY, 1999.

M. B. Mulder, K. L. Rauch, Sexual conflict in humans: Variations and solutions, Evolutionary Anthropology: Issues, News, and Reviews 18 (2009) 201–214. doi:10.1002/evan.20226.

R. Bird, Cooperation and conflict: The behavioral ecology of the sexual division of labor, Evolutionary Anthropology 8 (1999) 65–75. doi:10.1002/(SICI)1520-6505(1999)8:2<65::AID-EVAN5>3.0.CO;2-3.

R. B. Bird, B. F. Codding, The sexual divison of labor, in: R. Scott, S. Kosslyn (Eds.), Emerging Trends in the Social and Behavioral Sciences, 2015, pp. 1–16.

G. K. Hadfield, A coordination model of the sexual division of labor, Journal of Economic Behavior & Organization 40 (1999) 125–153. doi:10.1016/S0167-2681(99)00053-0.

N. M. Waguespack, The organization of male and female labor in foraging societies: Implications for early Paleoindian archaeology, American Anthropologist 107 (2005) 666–676. doi:10.1525/aa.2005.107.4.666.

J. Krause, G. D. Ruxton, Living in Groups, Oxford University Press, New York, NY, 2002.

Appendix A. Permanent Conflict

When the population only consists of **Stubborn 1** and **Stubborn 2**, the payoff of each strategy, conditional on i **Stubborn 1**s being present in a group, are given by:

$$\begin{aligned} V(S_1) &= \frac{1}{1-\omega}\left[b_1\frac{i+1}{n} + b_2\frac{n-i-1}{n} - c_1 - (n-i-1)(p_2^1+k_1^2)\right] \\ V(S_1) &= \frac{1}{1-\omega}\left[b_1\frac{i}{n} + b_2\frac{n-i}{n} - c_2 - i(p_1^2+k_2^1)\right] \end{aligned} \tag{A.1}$$

Each strategy's fitness function is given by:

$$\begin{aligned} W(S_1) = w_0 + \frac{1}{1-\omega}\Bigg[b_1\frac{E[i|S_1]+1}{n} + b_2\frac{n-E[i|S_1]-1}{n} - c_1 \\ -(n-E[i|S_1]-1)(p_2^1+k_1^2)\Bigg] \\ W(S_2) = w_0 + \frac{1}{1-\omega}\Bigg[b_1\frac{E[i|S_2]}{n} + b_2\frac{n-E[i|S_2]}{n} - c_2 \\ -E[i|S_2](p_1^2+k_2^1)\Bigg] \end{aligned} \tag{A.2}$$

Assuming there is no assortment of strategies then $E[i|\cdot] = (n-1)x$, where x is the frequency of **Stubborn 1**. **Stubborn 2** will increase relative to **Stubborn 1** so long as:

$$(c_1 - \frac{b_1}{n}) - (c_2 - \frac{b_2}{n}) + (n-1)\Big((1-x)[p_2^1+k_1^2] - x[p_1^2+k_2^1]\Big) > 0 \tag{A.3}$$

The stability criterion for $\mathbf{S_1}$ is found by setting $x = 1$ and flipping the greater-than sign. Alternatively, the stability criterion for $\mathbf{S_2}$ is given by setting $x = 0$.

Appendix B. Contrite $\mathbf{S_1}$

In this section we evaluate when a strategy that responds to $\mathbf{S_2}$ punishment by adopting behavior 2 has a higher fitness (and can come to replace) behavior 1 bearing strategies. Such a criterion is of interest because it determines the conditions under which a $\mathbf{S_2}$ mutant will be able to coerce cooperation with its preferred behavior (2). Without the criterion in this section being satisfied, no normative chance can occur within the existence of a focal group, i.e. no focal agent would experience normative change within the lifetime of its group.

Assume a basal population that consists of some mix of predominantly $\mathbf{S_1}$ and a small proportion of a variant of $\mathbf{S_1}$, which we label $\mathbf{S_{1\to 2}}$. $\mathbf{S_{1\to 2}}$ initially plays 1 and punishes any agent that fails to play 1, however, should any 2 playing agent punish them, they immediately shirk 1 and switch to contributing to 2, also punishing any agent that fails to contribute to 2. In other words, in groups where no $\mathbf{S_2}$-like phenotype exists, $\mathbf{S_{1\to 2}}$ looks exactly like a $\mathbf{S_1}$ and no relative fitness differences exist between the two strategies; in groups where any agent punishes behavior 1 contributors, $\mathbf{S_{1\to 2}}$ comes to adopt a behavior 2 playing phenotype.

Additionally, assume that both of these strategies have some probability of "mutating" and playing $\mathbf{S_2}$, which occurs with probability ϵ due to some form of non best-response play. We make the conservative assumption that no two agents in a group ever simultaneously adopt a $\mathbf{S_2}$ phenotype. Such non best-response play may reflect behavioral errors; agents failing to understand their environments, should they possess cognitive capacities; or, relatedly, agents probing their environments, perhaps because they rarely observe punishment in equilibrium.

Because fitness differences only exist when a $\mathbf{S_2}$ mistake occurs, we only examine the fitness differences in those groups:

$$W(S_1) : w_0 + b_1\frac{n-1}{n} + b_2\frac{1}{n} - c_1 - (p_2^1 + k_1^2) \\ +\frac{\omega}{1-\omega}\left[b_1\frac{n-1}{n} + b_2\frac{1}{n} - c_1 - (p_2^1 + k_1^2)\right] \tag{B.1}$$

$$W(S_{1\to 2}) : w_0 + b_1\frac{n-1}{n} + b_2\frac{1}{n} - c_1 - (p_2^1 + k_1^2) \\ +\frac{\omega}{1-\omega}\left[b_1\frac{n-2}{n} + b_2\frac{2}{n} - c_2 - (n-2)(p_1^2 + k_2^1)\right] \tag{B.2}$$

$\mathbf{S_{1\to 2}}$ has a higher fitness than $\mathbf{S_1}$ when:

$$(c_1 - \frac{b_1}{n}) - (c_2 - \frac{b_2}{n}) \\ > \\ (n-2)(p_1^2 + k_2^1) - (p_2^1 + k_1^2) \tag{B.3}$$

Appendix C. $\mathbf{S_{1\to 2}}$ and $\mathbf{S_2}$ Competition

Here we study a system whereby normative change can occur within the existence of a group. The ecology of strategies in this section assumes that Proposition 2 holds and that **Stubborn 2** is able to coerce cooperation with its preferred behavior (2).

Let i denote the number of $\mathbf{S_{1\to 2}}$ in a group besides the focal individual. The expected fitness for each strategy are given by:

$$W(S_2) = w_0 + b_1\frac{E[i|S_2]}{n} + b_2\frac{n - E[i|S_2]}{n} - c_2 - E[i|S_2](p_1^2 + k_2^1) \\ + \frac{\omega}{1-\omega}(b_2 - c_2) \tag{C.1}$$

$$W(S_{1\to 2}) = w_0 + b_1 \frac{E[i|S_{1\to 2}]+1}{n} + b_2 \frac{n - E[i|S_{1\to 2}] - 1}{n} - c_1 \\ - (n - E[i|S_{1\to 2}] - 1)(p_2^1 + k_1^2) \\ + \frac{\omega}{1-\omega}\left(b_2 - c_2 + Pr(i = n-1|S_{1\to 2})\left[(b_1 - c_1) - (b_2 - c_2)\right]\right) \tag{C.2}$$

We assume no assortment of strategies. Correspondingly, the expected number of $\mathbf{S_{1\to 2}}$ does not depend on a focal agent's strategy and $E[i|\cdot] = (n-1)x$, where x denotes the proportion of $\mathbf{S_{1\to 2}}$ in the population. Similarly, regardless of her strategy, the probability of a focal agent encountering $n-1$ $\mathbf{S_{1\to 2}}$s, $Pr(i = n-1|\cdot)$, is given by x^{n-1}.

$\mathbf{S_2}$ will increase in the population so long as:

$$(c_1 - \frac{b_1}{n}) - (c_2 - \frac{b_2}{n}) + (n-1)\left[(1-x)(p_2^1 + k_1^2) - x(p_1^2 + k_2^1)\right] \\ + \frac{\omega}{1-\omega}\left[(b_2 - c_2) - (b_1 - c_1)\right] x^{n-1} > 0 \tag{C.3}$$

This function is a convex function of x on the relevant interval for which it is defined, $[0,1]$ so long as $(b_2 - c_2) - (b_1 - c_1)$ is positive and behavior 2 confers larger net-benefits to the group than behavior 1. The all-$\mathbf{S_{1\to 2}}$ stable equilibrium (or invasion criterion for $\mathbf{S_2}$) is found by setting $x = 1$ and reversing the greater-than sign; the all $\mathbf{S_2}$ stable equilibrium is found by setting $x = 0$; the internal equilibrium cannot be solved for in closed form expression, but exists whenever neither of the strategies are evolutionarily stable. Should $\mathbf{S_{1\to 2}}$ be unstable but $\mathbf{S_2}$ be stable, two internal equilibria exist, an unstable one located

near $x = 0$ and a stable equilibrium near $x = 1$. This occurs so long as:

$$(n-2)(p_2^1 + k_1^2 + p_1^2 + k_2^1)\frac{(p_2^1 + k_1^2 + p_1^2 + k_2^1)(1-\omega)}{\omega\left[(b_2 - c_2) - (b_1 - c_1)\right]}^{1/(n-2)} \\ > \\ (p_2^1 + k_1^2)(n-1) + (c_1 - \frac{b_1}{n}) - (c_2 - \frac{b_2}{n}) \tag{C.4}$$

Should $(b_2 - c_2) - (b_1 - c_1)$ be negative and behavior 2 confers smaller net-benefits to the group than behavior 1, the fitness difference between the two strategies is a concave function of x on the relevant interval and the sign of the above criterion is flipped and the paired interior equilibrium occurs when $\mathbf{S_2}$ is unstable but $\mathbf{S_{1\to 2}}$ is stable. Now, the interior stable equilibrium occurs near $x = 0$ and the stable one near $x = 1$.