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 riddled with norms, many though not all of which are costly for individuals to adopt. Similarly, human ecological adaptation relies on costly-behaviors that often generate non-rivalrous and non-exclusionary benefits for group-members. Yet, in a dynamic world, innovations, environmental change, and information-revelation mean that what norms are beneficial for a group to adopt will inevitably change over time. However, multiple game-theoretic models studying the various mechanisms stabilizing normative behaviors have demonstrated that the stability of a norm does not depend on the benefits it confers. In turn, explanations of normative change have either relied on group-selective mechanisms to explain the presence of adaptive norms or have failed to identify conditions under which normative change occurs. We study normative change by means of costly-punishment and conflict resolution. We identify social differentiation in goals and punishment capacity as a key condition permitting normative change. While normative change that results from such social differentiation need not be group beneficial it will be beneficial to some subset of agents in the population. We additionally discuss how the intra-societal forces of normative conflict that we study might interact with group-selective forces and in turn determine the dynamics and outcomes of group-selection.

1. Introduction

Human life, compared to that of all other organisms, is uniquely structured by norms (Bicchieri, 2006; Boyd, 2019; Roughley and Bayertz, 2019; Searle, 1995). While much debate exists over how humans evolved to be normative and what a proper definition for norms is, scientists from diverse fields such as

[*] Corresponding author

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

anthropology and evolutionary theory (Boyd, 2019) to economics (Nunn, 2020) emphasize the role of norms and packages of norms, or institutions (Ensminger and Henrich, 2014; Henrich and Ensminger, 2014; Henrich and Muthukrishna, 2021), for determining human psychological, behavioral, and group-level outcomes.

Costly norms are behaviors whose adoption requires individuals to shoulder private-costs larger than the private-benefits they confer 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), and sanctions, either costly (Boyd and Richerson, 1992; Boyd et al., 2010) or extractive (Bhui et al., 2019), in addition to non-punishment mechanisms, such as costly signaling (Gintis et al., 2001), can stabilize such behaviors. Importantly, these mechanisms can stabilize behaviors regardless of their associated benefits. This leaves social and biological scientists in somewhat of a quandary because human groups are characterized by a plethora of adaptive and beneficial practices, technologies, norms, and institutions (Boyd, 2019; Kelly, 1995). While individuals might only seek to stabilize behaviors that are beneficial to them (Richerson and Henrich, 2012), individuals often cannot (or, in some cases, it is impossible for them to be able to) explain said benefits (Boyd et al., 2013; Henrich, 2002, 2016).

Something of a trade off also immediately becomes apparent in this story. It will often be possible to improve upon practices and norms should they be stabilized irrespective of their benefits. More so, human environments are also dynamic meaning that once adaptive behaviors can quickly become maladaptive and harmful to individuals and groups as the environment changes. In other words, individuals will often benefit from changing the norms that characterize their groups. Yet, if normative change occurs too easily, the social world may come to be chaotic and unpredictable for individuals and norms that have causally opaque benefits (Boyd, 2019; Boyd et al., 2013; Derex et al., 2019; Henrich, 2016) or, in the language of reinforcement learning, offer sparse rewards (Köster et al., 2022), would likely be abandoned, much to individuals' detriment. At the same time, if normative change can never occur, then adaptive changes and innovations will never be adopted, again, harming group-members.

Consider a scenario whereby some drift-like procedure and stochastic sampling has produced a segregated population of individuals for whom eating red-berries violates a norm and correspondingly group-members punish norm-violators by means of either costly-sanctions (Boyd and Richerson, 1992) or refusing them the benefits of mutual aid (Panchanathan and Boyd, 2004). Should an individual within the group discover, through their private and illicit experimentation, that red berries are actually quite healthful and calorie dense, such an innovation would never be adopted by the broader community. Any individual advocating that red berries should be eaten will quickly come to be characterized by lower payoffs than those agents who conform to the norm once sanctioned assuming that sanctions are sizable. Panchanathan and Boyd (2004) have demonstrated that in a population where a costly-norm has been stabilized by strategies that withhold (beneficial) mutual aid from norm-breakers, an alternative costly-norm cannot invade even when said norm provides greater benefits *to all* group-members and all individuals would in fact prefer that the group adopts the alternative norm: "Differing opinions of good citizenship and impropriety have driven a moral wedge into the community" (Panchanathan and Boyd, 2004, p.501). The problem Panchanthan and Boyd identify is not a new one. Locke, in offering decentralized and norm-based punishment as an alternative to Hobbes' centralized Leviathan, emphasized the potential for conflict among factions that prefer distinct norms to in turn disrupt social-order

(Barrett, 2020). Assuming that individuals adopt higher-payoff strategies, then in order for the red-berry convert to alter its group's norm, it would need to engage in and win normative conflict with the resident strategies, punishing those who do not eat red berries.

Such heterogeneous normative values may emerge through endogenous cultural evolutionary processes themselves. Paddy rice farming can produce extremely large caloric surpluses when the proper labor inputs are supplied (Bray, 1986). Such returns facilitate the diversification of economic roles; the historical commercialization and monetization of southern China likely depended on paddy rice's unique caloric surplus (Noblit, 2021). To produce such large surpluses, paddy rice agriculture requires extensive cooperation and coordination in terms of a community's planting practices, water control, and irrigation infrastructure maintenance (Bray, 1986). These cooperative behaviors necessitate mechanisms to ensure agents do not free-ride on the efforts of others (Boyd and Richerson, 1992). Villages that are able to stabilize cooperative costly-norms concerning paddy rice growing activities will produce large surpluses and correspondingly will be able to economically diversify and specialize. Yet, as individuals adopt novel economic roles, such as if some subset of the community becomes foresters, they will also likely come to prefer their community adopt distinct norms that more directly reflect their endogenous preferences, for example the costly enforcement of forest commons management (Rustagi et al., 2010) or preserving forest-land for forestry purposes as opposed to converting it into more paddy land (Bray, 1986).

Cultural evolutionists have largely sought to explain the question of normative change and the generally adaptive nature of costly-norms and institutions by means of cultural group selection (CGS) whereby equilibria (stabilized norms and institutions) proliferate proportional to their associated payoffs (Boyd and

Richerson, 2009, 2002, 2009; Henrich, 2004; Richerson et al., 2016). Generally, cultural group selection is conceptualized to act at the level of groups, such as when different norms and institutions lead groups to go extinct at variable rates or heterogeneously succeed in inter-group combat; in such cases, normative change occurs only in an aggregate distribution of equilibria and no intra-societal normative change need occur. Theorists have also pointed to important group-selective forces that may help explain intra-societal adaptive normative change, such as when individuals imitate the behaviors from nearby groups proportional to their payoff differences (Boyd and Richerson, 2002).

However, normative change regularly occurs because of intra-societal forces. Though such processes are often difficult for ethnographers and historians to capture, researchers have highlighted the political, legal, and generally collective processes that determine intra-societal normative establishment and change (Boehm et al., 1996; Cohen and Middleton, 1967; Hadfield and Weingast, 2013). More so, in both the red-berry and forestry scenarios we described above, as well as in Locke and Panchanthanan and Boyd's formulation of normative conflict, individuals clearly vary in terms of what norms they wish to see characterize their groups. Normative conflict may erupt around normative goal-seeking and such conflict has been a source of both pro-social (Keyssar, 2009) and regressive or anti-social normative change (Bello, 2019), playing a key role in social evolution during the Holocene of the past $\approx 12,000$ years. For example, shifting backwards in time, it has characterized the Leveller and Digger movements in 17th century England (Rees, 2017; Wood, 2017), the Guild movements of medieval Europe (Ogilvie, 2021), efforts by the Roman plebs seeking release from debt obligations to the patricians and more well-established legal rights (Forsythe, 2005), and even collective labor action by Egyptian scribes in 12th century BCE (Edgerton, 1951). The ethnographic political and legal anthropological literature is rife with cases of agents who possess a normative, either group-beneficial or -harmful, vision for their communities; importantly, said individuals often seek to coerce group-members' cooperation with and conformity to said vision (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).

One might assume that normative conflict only characterizes the relatively more extreme social differentiation that has occurred during the Holocene and that prior to said differentiation, people were generally homogeneous in their preferences. However, as our red-berry case exemplifies, individuals may seek distinct goals because of what private information they possess. More so, preference-heterogeneity is part and parcel of evolution generally and hominin evolution specifically, arising from 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 human 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 (Bird, 1999; Bird and Codding, 2015; Hadfield, 1999; Waguespack, 2005). Such variation may lead individuals to correspondingly prefer different collective movement decisions, which are of obvious relevance to foragers and social organisms generally (Krause and Ruxton, 2002), group-compositions, or distributions of individuals' time-allocations at the sub-household level. 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 learning. That such differences exist and produce normative conflict is central to a major theory of human evolution. In chimpanzee dominance hierarchies, a single male or small-coalition of males seeks their private reproductive interests by imposing behaviors on group-members whereas human foragers are generally characterized by norms and practices that limit such competition, hierarchy, and exploitation, often responding violently to self-aggrandizing and bullying individuals who seek to elevate themselves and their interests above others (Boehm et al., 1993; Boehm, 1999). Long-standing ethnographic and anthropological work refutes a simplistic notion of an egalitarian human past (Buitron and Steinmüller, 2021; Lévi-Strauss, 1945; Lowie, 1948; Mauss, 2004; Sahlins, 2017; Wengrow and Graeber, 2015), necessitating an understanding of normative conflict resolution: When does it produce prosocial and group-beneficial ends; when does it benefit only a few individuals in a community and confer costs to others (Boehm, 1999; North et al., 2009)?

We argue that intra-societal processes producing and resolving normative conflict are important for understanding normative change and necessitate direct study. These processes may even permit adaptive and prosocial normative change. While cultural evolutionary work has largely relied on group-selective mechanisms to explain (adaptive) normative change, other anthropologists have argued that the empirical distribution of norms instead largely reflects bargaining processes defined by the distributional consequences of alternative possible rules for individuals (Ensminger and Knight, 1997; Knight, 1992). While, strictly speaking, these authors fail to move beyond dyadic interactions and thus cannot speak robustly of norms that are stabilized by third parties, their argument that the empirical distribution of norms reflects individual goal-seeking appears to contradict the consistent theoretical finding that costly-norms are stabilized regardless of their benefits (Boyd and Richerson, 1992; Gintis et al., 2001; Panchanathan and Boyd, 2004). We do not consider these differences to be irreconcilable, but rather consider them to study the same issue, norms and norm-stabilization, from different assumptions and vantage points using highly simplified models. Cultural evolutionary results explain why arbitrary behaviors can be stabilized, but make no claims concerning the basis on which individuals choose to act; strategies are hard coded and are not chosen, strictly speaking. Ensminger and Knight contrarily assume that strategies are chosen through Nash-like reasoning processes requiring that a focal individual, in some sense, understands the (epistemically demanding) primitives of a game, i.e. who is playing the game, each player's action set, what outcomes are possible, the various payoffs each player experiences for each outcome, etc. A cultural evolutionist would refer to such an explanation as a proximate one because it ignores and cannot speak to the questions a cultural evolutionist centralizes: how preferences, institutions and rules, beliefs, and even players originate and spread.

More so, cultural evolutionary models, because they make no claims concerning choice processes, can accommodate different ways of choosing, i.e. these models reflect that (1) humans adopt strategies and behaviors through various mechanisms (Kendal et al., 2018; Poulin-Dubois and Brosseau-Liard, 2016; Rendell et al., 2011) and, relatedly, that (2) choosers exist in a world defined by Knightian uncertainty (Knight, 1921). When a chooser is a naive child, a migrant to a new community or region, presented with a novel and unfamiliar social context, or when a chooser lacks causal knowledge about the world, agents will lack the information required to choose in ways modeled by classical game theory. Correspondingly, punishers may be willing punish on the basis of unfalsifiable beliefs or even incorrect beliefs about causal processes. As any student of econometrics or public-health can attest, reliably inferring causal knowledge is no small feat. If agents develop beliefs, for exogenous or endogenous reasons, that visiting certain states will produce negative rewards, learning processes may prevent them from exploring such states (LeDoux et al., 2017). Cultural evolutionary game-theoretic models can accommodate strategies built around such beliefs, which are of particular interests when such punishment strategies are built around beliefs, which may spread according to their functional effects because of adaptive social learning strategies (Henrich, 2016), particularly when they entail externalities, i.e. the frequent ethnographic finding that one's failure to adopt or avoid a behavior will produce negative beliefs for group-members (Henrich, 2016; Sahlins, 2017), motivating group-members' punishment.

Importantly, even in controlled lab experiments where study-participants need track relatively few variables and deal with far less uncertainty when compared to *in situ* environments, individuals cannot anticipate the consequences of simple changes to the simplest of institutions (Gurerk et al., 2006). This is surprising considering that the studied institutions reflect the game-theoretic structures that humans have depended on for at least tens of thousands of years (Boyd and Richerson, 2022). Such empirical findings suggest that our ingrained ability to reason correctly about or predict the causal impact of altering norms and institutions is minimal. Importantly, a lack of reliance on causal insight and individual learning as opposed to social learning biases may be exactly what permits human populations to accumulate complex technologies and institutions (Boyd, 2019; Boyd et al., 2013; Henrich, 2016). How individuals reason and choose is a function of genetic and cultural evolutionary processes; even individual-learning is conditioned on said processes and occurs within environments constructed by them (Henrich and Muthukrishna, 2021; Heyes, 2019; Muthukrishna et al., 2021). Because Ensminger and Knight consider players to choose under conditions of Savagean risk (Savage, 1954) rather than uncertainty, a requisite assumption to permit rational-choice modeling, they necessarily restrict their class of explanations to proximate descriptions, ignoring the endogenous nature of preferences and psychology (the mental processes impacting choice) within cultural and genetic evolutionary processes. That being said, of course Ensminger and Knight are obviously correct to argue that individuals differ from one another in ways that relate both to what their (endogenized) goals are and their abilities to enforce those goals as well as the fact that individuals will act to pursue their goals when they can. We seek to integrate this perspective into a cultural evolutionary account of normative change.

To study the role of normative change in human social evolution, we construct a stylized model that permits individuals to both prefer distinct behaviors, to engage in conflict, and for conflict to be settled and normative change to occur. Following Boyd and Richerson (1992) we model conflict as the decentralized allocation of punishment-costs at cost to the agent. We do this not because we think this form of costly-punishment provides a sufficient description for how all costly-norms are stabilized. This is known to be false. Various mechanisms can stabilize costly-behaviors in n-player settings (Bhui et al., 2019; Boyd et al., 2010; Gintis et al., 2001; Henrich et al., 2015; Panchanathan and Boyd, 2004). Empirical evidence confirms the relevance of a multiplicity of costly-norm enforcement mechanisms and demonstrates that societies vary in whether and how punishment is used in both third-party cooperative (Henrich et al., 2006) and public-good (Herrmann et al., 2008) settings. Instead, we study decentralized forms of sanctioning because it provides an analytically tractable manner for examining the role of social-differentiation on the evolution of norms while simultaneously capturing the capacity for individuals to allocate costs on one another and for this conflict to be resolved. Our model provides a baseline for comparison with future and alternative models and not a single theory of normative change.

In order to construct a sufficiently general model of conflict resolution, we model its occurrence two ways. In the first model we present, akin to Boyd and Richerson's (1992)) seminal work, we assume no political or legal institutions exist to resolve normative conflict. Instead, we ask when resident strategies' best-response to the experience of punishment is to abandon their current norm and adopt an alternative norm, permitting normative change. Our second model recognizes that societies often are characterized by institutionalized forms of conflict resolution. We correspondingly model include a mechanism, $\boldsymbol{\tau}$ that settles conflict stochastically in favor of either norm or permits conflict to continue.

We show that permitting normative conflict means that the stability criterion of a norm no longer excludes the benefit associated with the behavior. Instead, when social differentiation has produced agents defined by heterogeneous preferences and punishment capacities, then agents can coerce normative change if the cost of conflict that they impose on others exceeds the cost of conflict imposed on them. If such a criterion holds then a norm is stable only if the benefit to the mutant type in switching its group to its preferred norm is less than the cost of conflict that the mutant experiences. A similar criterion holds when we consider a mechanism that stochastically determines what norm a group adopts after a single round of normative conflict has occurred. We take these results to indicate that the determinants of how novel behaviors produce heterogeneity with respect to (1) how agents experience the punishment efforts of others and (2) their ability to confer punishment on others have a significant role in determining normative change. Finally, we discuss how our findings may interact with cultural group selection to determine its dynamics and results.

2. Model

2.1. Framework

Our model follows standard assumptions of evolutionary game-theoretic models examining the role of punishment in stabilizing public-good contributions (Boyd and Richerson, 1992). Atypically, we permit two types, labeled Red and Blue, to exist in an effectively infinitely sized meta-population. Each type is drawn from separate and extremely large sub-populations. Adopting this formalism permits us to study cases where strategies receive heterogeneous payoffs from the same outcomes while maintaining a cultural evolutionary game-theoretic framework. Strategies are sampled according to their frequency into groups of size n where m strategies are drawn from the Red sub-population and $n - m$ strategies from the Blue population; m need not equal $\frac{n}{2}$. Once sampled into groups, strategies play two stage-games, described below, after which each group continues to exist with probability ω. To facilitate analytical tractability, we assume that no implementation errors occur and that results concerning implementation errors carry over from previous work (Boyd and Richerson, 1992).

The first stage-game is an n-player linear public goods game (Boyd and Richerson, 1992; Taylor, 1987). Agents can cooperate (contribute) or defect (free-ride) on either but not both of two behaviors: 1 or 2. The benefit of each public-good for a given agent depends on the agent's type. Red agents prefer behavior 1, meaning that $b_R^1 > b_R^2$, whereas Blues prefer behavior 2 such that $b_B^2 > b_B^1$. That both behaviors are public goods means that if a Blue agent adopts behavior 1 then she necessarily confers $\frac{b_R^1}{n}$ to every Red agent and $\frac{b_B^1}{n}$ to every Blue agent in her group. Finally, when agents adopt a public-good behavior, they shoulder type-specific private costs defined as c_i^j where again i, j index the focal agent's type and the public-good behavior the agent adopted respectively.

Our asymmetric formalism aims to model situations when the members of a society can be separated into sub-populations conditional on how they are impacted by the various involved behaviors. Correspondingly, agents may "prefer" distinct norms conditional on their type. Such heterogeneous "preferences" may exist for numerous biological and social reasons. Group-members may differ in (1) their role in sexual reproduction, such that they experience different costs and benefits for a given behavior; (2) their economic traits or characteristics, such as being tenants or landowners or relying more heavily on farming or pastoralism; or (3) their social identities may entail distinct and previously stabilized rights and responsibilities, such as might results from economic specialization and a division of labor(Henrich and Boyd, 2008) or the presence of a chiefly clan that maintains rights over the preferential use of violence or monopoly over land or key resources (Earle and Spriggs, 2015; Kirch, 2010). Critically, we assume that these types are exogenously fixed and that a member of one sub-population cannot ever become the member of another sub-population.

The second stage-game is a punishment game whereby agents can adopt a cost, $k_{i,o}^{j,p}$, to punish another agent $p_{i,o}^{j,u}$ units where i and and o index the type and adopted behavior, 1 or 2, of the focal individual and j and u index the type adopted behavior of their social partner. For instance, if a Blue 1 agent punishes a Red 2 agent, the focal Blue agent adopts a cost $k_{B,1}^{R,2}$ to impose $p_{B,1}^{R,2}$ units cost on the Red agent. We index on an actor and her social partner's type and behaviors to formally investigate how social differentiation impacts the evolution of norms when individuals have the ability to impose heterogeneous punishment costs compared to other strategies in the group. Such social differentiation, particularly in how agents can impose costs on one another, is likely critical to understanding the evolution of human societies (North et al., 2009) and need not rely on the cultural evolution of and reliance on agricultural technologies to develop (Arnold et al., 2016).

Importantly, our punishment capacities are not only type-dependent but also depend on what norm the agent is characterized by. This is a somewhat strange formalism but necessary in light of higher-order punishment concerns (Boyd and Richerson, 1992). For instance, a Blue strategy playing norm 1, $\mathbf{B_1}$, in conflict with a $\mathbf{R_2}$ strategy, experiences a cost of $k_{B,1}^{R,2}$ units and inflicts $p_{B,1}^{R,2}$ units cost. However, should $\mathbf{B_1}$ be coerced to play norm 2 in the second and all future rounds (we later label this strategy $\mathbf{B_{1\to 2}}$), its punishment parameters will become that of a $\mathbf{B_2}$ strategy after which, when in conflict with a $\mathbf{R_1}$ strategy, it will experience a cost of $k_{B,2}^{R,1}$ units to inflict $p_{B,2}^{R,1}$ units punishment. In effect, we are modeling a process whereby what norm an agent plays interacts with its type to define their various punishment parameters.

The existence of conflict between two parties means that conflict can either continue indefinitely, resulting in ongoing rounds of simultaneous punishment among various agents, or that it may be resolved. We construct two models to consider two types of conflict resolution. The first model assumes that no institutional conflict resolution mechanism exists and that population-members have no rules concerning norm establishment and change (Hart, 1994). In such a case, normative change will occur only if a subset of actors can coerce cooperation from another subset. Correspondingly, we derive the conditions under which a strategy is able to coerce other strategies' acquiescence or cooperation. Conflict resolution in this case occurs deterministically when we assume the criterion defining such a best-response holds. The second model assumes some conflict resolution mechanism exists, labeled $\boldsymbol{\tau}$. This mechanism can make three possible rulings whenever simultaneous punishment occurs: with probability τ_1 it rules that the group should adopt 1, with probability τ_2 that norm 2 should be adopted, and with probability $1-(\tau_1+\tau_2)$ that conflict may continue. We collect these three probabilities in the vector $\boldsymbol{\tau} = [\tau_1, \tau_2, 1 - (\tau_1 + \tau_2)]$. These probabilities are fixed and exogenously determined.

We do not derive the conditions under which conforming to $\boldsymbol{\tau}$'s ruling is a best-response. Instead, following previous work examining incentives to conform to the normative rulings of third-parties (Hadfield and Weingast, 2012), we assume that an institution has previously been stabilized in the population such that individuals are incentivized to conform to its rulings. We conceptualize this as follows: While agents may disagree with what behaviors they would prefer their groups adapt, i.e. how behaviors are normatively labeled by the community, they still conform to what Hart (1994) refers to as secondary rules or the meta-rules defining how rules are made and changed in the community. Correspondingly, all agents, regardless of type, do agree that they will punish any agent that does not conform to $\boldsymbol{\tau}$'s ruling (Boyd and Richerson, 1992; Hadfield and Weingast, 2012). We do not adopt this formalism to provide a conclusive model of conflict resolution; no single model could likely accommodate all cross-cultural variation in how conflict resolution occurs. Instead, we seek to study how the presence of an institutions producing conflict resolution, that individuals will conform to, impacts normative change. Future modeling work must characterize distinct conflict resolution mechanisms and model them as players themselves.

2.2. Strategies

2.2.1. Deterministic Conflict Resolution

We initially define two strategies within each sub-population that are symmetric. The Red population is defined by two strategies: the first, labeled $\mathbf{R_1}$, always adopts behavior 1 and punishes any agent that does not adopt 1; $\mathbf{R_2}$, as one might expect, is a mutant Red strategy that always adopts 2, its non-preferred behavior, and punishes any agent that plays 1. The Blue population similarly is defined by two strategies: $\mathbf{B_1}$, which always adopts 1 and punishes any agent that does not cooperate by adopting 1 and $\mathbf{B_2}$, which always adopts 2 and punishes any agent that does not cooperate by adopting 2. Though we restrict our analysis to a static evolutionary analysis (Broom and Rychtář, 2013), we assume that mixed-equlibria where strategies initially defect before adopting either norm are irrelevant (Boyd and Richerson, 1992; Boyd et al., 2010). Similarly, we assume that no mixed-equilibria of norms exist, and, importantly (Taylor, 1979), that a mutant will never spontaneously play an non-preferred norm. Correspondingly, we assume that an invasion analysis need only study $\mathbf{B_2}$ and $\mathbf{R_1}$ mutants. The invasion of $\mathbf{R_2}$ and $\mathbf{B_1}$ mutants is assumed to only occur as the meta-population shifts norms.

2.2.2. Stochastic Conflict Resolution

In our second model where we implement the mechanism $\boldsymbol{\tau}$ that, following the occurrence of conflict, makes a ruling that group-members should either adopt either norm or that conflict may continue. Strategies in this stochastic conflict resolution model mirror those in the deterministic case with the exception that they are assumed to conform to $\boldsymbol{\tau}$'s ruling concerning what public good they play from the second round forward. Strategies additionally no longer punish from the second round forward. For example, in groups where $\mathbf{R_1}$ and $\mathbf{B_2}$ meet, conflict will occur. Once conflict occurs, $\boldsymbol{\tau}$ will make a ruling. In cases where $\boldsymbol{\tau}$ rules that the group should adopt norm 1, all agents, Red and Blue, will adopt norm 1 from the second round forwards and will cease engaging in punishment. We do not explicitly model $\boldsymbol{\tau}$ as a strategy but as an exogenous mechanism.

3. Results

3.1. A Norm's Stability

3.1.1. Deterministic Conflict Resolution

Considering the importance of punishment for costly-norm enforcement (Boyd and Richerson, 1992; Boyd et al., 2010; Mathew, 2017; Mathew and Boyd, 2011) and the potential for multiple punishing coalitions to emerge in a society and engage in conflict (Barrett, 2020), we are interested in understanding the conditions under which a meta-population characterized by a previously stabilized norm, 1, is stable to the invasion of a novel behavior and potential norm, 2. We conceptualize this process as follows: Norm 1 has previously been stabilized in a meta-population even though Blue agents categorically prefer norm 2. At some point, behavior 2 becomes possible and a mutant Blue agent, labeled $\mathbf{B_2}$, arises that not only plays 2 but also attempts to stabilize this norm through punishment. As previously discussed, we do not permit the simultaneous mutation of both $\mathbf{B_2}$ and $\mathbf{R_2}$ mutants (Taylor, 1979) under the assumption that a Red agent would never willingly play or attempt to enforce norm 2 unless Red agents experience conflict with $\mathbf{B_2}$s sufficiently often in expectation.

We first examine the stability criterion for norm 1 when there is no conflict resolution and no strategies concede, i.e. punishment continues indefinitely. The Blue mutant, $\mathbf{B_2}$, plays norm 2 and is punished once its play is observed by group members; observing its $n-1$ group members play 1, the mutant punishes each of them. Because none of the strategies are encoded to respond to punishment, this occurs for each and every round. Meanwhile, the average resident Blue strategy, $\mathbf{B_1}$, experiences no punishment. Under such conditions, $\mathbf{B_2}$ can never invade the resident Blue population and no norm change can occur:

Proposition 1. *(proof in Appendix A) If conflict among norm* 1 *and* 2 *players is never resolved but continues indefinitely, then norm* 1 *($\boldsymbol{B_1}$) is stable to an* *invasion by norm 2 ($\boldsymbol{B_2}$) so long as the cost to $\boldsymbol{B_2}$ for engaging in conflict with the $n-1$ norm 1 playing resident group-members exceeds any potential cost-savings that $\boldsymbol{B_2}$ experiences when adopting its preferred norm:*

$$\overbrace{(c_B^1 - \frac{b_B^1}{n}) - (c_B^2 - \frac{b_B^2}{n})}^{\textit{Potential cost-savings to } \boldsymbol{B_2} \textit{ from adopting norm } 2}$$

$$<$$

$$\underbrace{(p_{R,1}^{B,2} + k_{B,2}^{R,1})m + (p_{B,1}^{B,2} + k_{B,2}^{B,1})(n-m-1)}_{\textit{Cost to } \boldsymbol{B_2} \textit{ of punishing and being punished by } n-1 \textit{ resident strategies}}$$

This criterion states that $\mathbf{B_2}$ can invade only if its cost of punishing (the k terms) and being punished by (the p terms) the resident types is offset by the preferred behavior being less-costly to adopt, $(c_B^1 - \frac{b_B^1}{n}) - (c_B^2 - \frac{b_B^2}{n})$. While it is entirely reasonable to think that the Blue types might prefer behavior 2 because it is less costly for them to adopt, it is unrealistic in light of ethnographic evidence (Boehm, 1999) to assume that the change in net-private costs from adopting the preferred behavior could offset the cost of conflict with the resident strategies. Correspondingly, we would never expect norm 2 to spread through the population. Note, importantly, that the punishment the mutant imposes on resident types is not included in the criterion; the average resident Blue strategy will not encounter any mutants and correspondingly, the mutant $\mathbf{B_2}$ agent's punishment capacity is irrelevant.

This result however is highly unintuitive in comparison to historical data. It suggests that a norm that is stabilized by punishment can never be invaded by an alternative norm, regardless of whether the mutant is able to impose massive costs on all group members. Such a result is of course ridiculous from an empirical position, but also a theoretical one. In the former case, even a cursory reading of political history indicates that violence is regularly invested in to coerce behavioral change. However, from a theoretical position, one of the assumptions required to derive Proposition 1 is that conflict continues indefinitely and is never resolved in any party's favor. Two forces suggest that we should not expect groups to be characterized by intensive ongoing conflict. First, cultural group selection (Boyd and Richerson, 2002, 2009; Richerson et al., 2016) mechanisms would select against such societies and second, and more importantly, strategies that avoid the costs of ongoing conflict will likely be fitter than strategies willing to continue to engage in conflict (Boyd and Richerson, 1992). We now define the conditions under which strategies that acquiesce to the mutant's punishment by adopting its preferred norm will be fitter (Boyd and Richerson, 1992).

Consider a resident Blue strategy that plays norm 1. Observing the costs associated with the initial round of conflict with $\mathbf{B_2}$, the resident strategy has the option of continuing to engage in conflict or adopting behavior 2 so that it is no longer a target of $\mathbf{B_2}$'s ire. We assume that $\mathbf{B_2}$ is committed to engaging in repeated conflict. We creatively label the mutant of $\mathbf{B_1}$ that acquiesces to $\mathbf{B_2}$ after observing the costs of engaging in conflict, $\mathbf{B_{1\to 2}}$. $\mathbf{B_1}$ should acquiesce to punishment so long as engaging in conflict with the mutant $\mathbf{B_2}$ strategy is costlier than engaging in conflict with the resident $n-2$ strategies:

Lemma 1. *(proof in Appendix B) If there is some probability of $\boldsymbol{B_2}$ arising in any group, then a strategy, $\boldsymbol{B_{1\to 2}}$, that adopts norm* 2 *and punishes norm* 1 *players after conflict with $\boldsymbol{B_2}$, will propagate at the expense of the resident recalcitrant strategy, $\boldsymbol{B_1}$, which does not respond to punishment, so long as a single $\boldsymbol{B_2}$ mutant is able to impose greater costs on $\boldsymbol{B_1}$ than the resident* *strategies can impose on* $\boldsymbol{B_{1\rightarrow 2}}$*:*

$$\overbrace{(p^{B,1}_{B,2} + k^{B,2}_{B,1})}^{\textit{B}_1\textit{'s cost of conflict with }\textit{B}_2} + \underbrace{(c^1_B - \frac{b^1_B}{n}) - (c^2_B - \frac{b^2_B}{n})}_{\textit{Private benefit to } B_{1\rightarrow 2} \textit{ in adopting 2}}$$

$$>$$

$$\underbrace{(p^{B,2}_{R,1} + k^{R,1}_{B,2})m + (p^{B,2}_{B,1} + k^{B,1}_{B,2})(n-m-2)}_{\textit{B}_{1\rightarrow 2}\textit{'s cost of conflict with resident strategies}}$$

A qualitatively identical Lemma and criterion exists for the Red agents:

Lemma 2. *(proof in Appendix B) If there is some probability of* $\boldsymbol{B_2}$ *arising in any group then a strategy,* $\boldsymbol{R_{1\rightarrow 2}}$*, that adopts norm* 2 *and punishes norm* 1 *players after it engages in conflict with* $\boldsymbol{B_2}$*, will propagate at the expense of the resident recalcitrant strategy,* $\boldsymbol{R_1}$*, which does not respond to punishment with cooperation, if a single* $\boldsymbol{B_2}$ *mutant is able to impose greater costs on* $\boldsymbol{R_1}$ *than the resident strategies can impose on* $\boldsymbol{R_{1\rightarrow 2}}$*:*

$$(p^{R,1}_{B,2} + k^{B,2}_{R,1}) + (c^1_R - \frac{b^1_R}{n}) - (c^2_R - \frac{b^2_R}{n})$$

$$>$$

$$(p^{R,2}_{R,1} + k^{R,1}_{R,2})(m-1) + (p^{R,2}_{B,1} + k^{B,1}_{R,2})(n-m-2)$$

When the criteria found in Lemmas 1 and 2 hold, resident strategies initially attempt to coerce cooperation from mutant $\mathbf{B_2}$s but, observing the costs of this conflict, best-respond by adopting norm 2 and avoiding further conflict with the mutant. Normative change is now possible and groups consisting of a single mutant $\mathbf{B_2}$ will now adopt norm 2 after the first round.

We can now establish our main result: the stability criterion for $\mathbf{B_{1\rightarrow 2}}$ and correspondingly for norm 1. If this criterion is met then norm 1 is stable to norm 2 invasion, otherwise $\mathbf{B_2}$ and norm 2 can invade the Blue population.

Proposition 2. *(proof in Appendix C.1) So long as* $\boldsymbol{B_2}$ *can coerce cooperation from resident Blue strategies but not resident Red strategies (i.e., Lemma 1 holds but Lemma 2 does not), then* $\boldsymbol{B_{1\to 2}}$ *(and norm* 1*) is stable to a* $\boldsymbol{B_2}$ *(and norm* 2*) invasion so long as the cost to* $\boldsymbol{B_2}$ *of a single round of conflict with* $\boldsymbol{B_{1\to 2}}$ *and ongoing conflict with the Red resident strategy,* $\boldsymbol{R_1}$*, exceeds the life-time benefit to* $\boldsymbol{B_2}$ *of stabilizing its preferred norm among the* $n-m$ *Blue group-members plus any private cost-savings that* $\boldsymbol{B_2}$ *enjoys in adopting norm* 2*:*

$$\overbrace{(p^{B,2}_{B,1}+k^{B,1}_{B,2})(n-m-1)}^{\textit{Cost to } \boldsymbol{B_2} \textit{ of conflict with } \boldsymbol{B_{1\to 2}} \textit{ for one round}}+\underbrace{\frac{(p^{B,2}_{R,1}+k^{R,1}_{B,2})m}{1-\omega}}_{\textit{Cost to } \boldsymbol{B_2} \textit{ of ongoing conflict with } \boldsymbol{R_1}}$$

$$>$$

$$(b^2_B-b^1_B)\frac{1}{n}+\underbrace{\frac{(c^1_B-c^2_B)+\omega(b^2_B-b^1_B)(1-\frac{m}{n})}{1-\omega}}_{\textit{Lifetime benefit to } \boldsymbol{B_2} \textit{ of coercing cooperation from Blue fraction of group}}$$

If we permit Lemma 2 to additionally hold and cooperating with $\mathbf{B_2}$ agents is a best-response for Red residents, then proposition 2 becomes somewhat simpler:

Proposition 3. *(proof identical to that found in Section 3.1 but additionally enforcing Lemma 2)) If the best-response for both Blue and Red resident strategies is to adopt norm* 2 *after engaging in conflict with a mutant* $\boldsymbol{B_2}$ *agent (i.e. both lemmas 1 and 2 hold), then* $\boldsymbol{B_{1\to 2}}$ *(and norm* 1*) is stable to a* $\boldsymbol{B_2}$ *(and norm* 2*) invasion so long as the cost to* $\boldsymbol{B_2}$ *of a single round of conflict with the resident strategies exceeds (1) any private-cost-savings it experiences in adopting norm* 2 *and (2) the benefit of switching its group from norm* 1 *to norm* 2*:*

$$(p^{B,2}_{B,1}+k^{B,1}_{B,2})(n-m-1)+(p^{B,2}_{R,1}+k^{R,1}_{B,2})m$$

$$>$$

$$(b^2_B-b^1_B)\frac{1}{n}+\frac{(c^1_B-c^2_B)+\omega(b^2_B-b^1_B)}{1-\omega}$$

Because both resident Red and Blue strategies now acquiesce to $\mathbf{B_2}$ punishment, $\mathbf{B_2}$ is able to coerce cooperation with norm 2 from all group-members in the groups it shows up in. Correspondingly, the stability criterion for a norm now entails the benefit associated, to the Blue type, i.e. the type of the agent able to coerce cooperation, with each behavior.

3.1.2. Stochastic Conflict Resolution

In the previous section we assumed that conflict was resolved deterministically and that a mutant strategy was able to coerce cooperation from group-members in all groups it shows up in. We now assume that instead a stable mechanism, $\boldsymbol{\tau}$, exists that settles conflict stochastically. More so, we assume that all agents are incentivized to coopearte with this mechanism, i.e. that historical processes stabilized behavioral adherence to the mechanism's ruling. This means that in groups where $\mathbf{B_1}$ mutants exist, all agents will adopt norm 2 starting from the second round, with probability τ_2.

Proposition 4. *(proof in section Appendix C.2) Norm* 1 *(and* $\boldsymbol{B_1}$*) is stable to a norm* 2 *(and* $\boldsymbol{B_2}$*) invasion so long as the probability of* $\boldsymbol{\tau}$ *making a norm* 2 *ruling times the change in benefits from adopting norm* 2 *to Blue agents, plus any private-cost benefit to Blue agents for adopting norm* 2 *is less than the cost to* $\boldsymbol{B_2}$ *of the first round of conflict plus any risk of ongoing future conflict:*

$$\omega \underbrace{\tau_2}_{\textit{Probability } \boldsymbol{\tau} \textit{ rules to adopt } 2} \overbrace{(b_B^2 - b_B^1)}^{\textit{Benefit of norm 2 compared to 1 for Blue agents}} + (1-\omega\tau_1)(c_B^1 - c_B^2) + [1-\omega(\tau_1+\tau_2)]\left(\frac{b_B^2 - b_B^1}{n}\right)$$

$$<$$

$$\underbrace{[1-\omega(\tau_1+\tau_2)]\left[(p_{B,1}^{B,2} + k_{B,2}^{B,1})(n-m-1) + (p_{R,1}^{B,2} + k_{B,2}^{R,1})m\right]}_{\textit{Cost to } B_2 \textit{ of a single round of conflict + possible future rounds}}$$

Again, we see that the stability of a norm depends on the a difference in benefits of relative norms. More so, this mechanistic model demonstrates that this difference in benefits enters into the stability criterion *to the degree* that normative change occurs (τ_2).

3.2. A New Norm's Stability

3.2.1. Deterministic Conflict Resolution

In Section 3.1.1, we examined when a previously stabilized norm was stable to the invasion of a novel norm when proponents of the novel norm attempt to coerce cooperation by means of punishment. We now ask the conditions under which a meta-population defined by the novel norm, 2, is stable to a re-invasion by norm 1 adhering strategies, $\mathbf{B_{1\to 2}}$ and $\mathbf{R_{1\to 2}}$. We again assume that mutant strategies do not spontaneously play a not-preferred norm and need only examine the invasion process in the Red sub-population.

Importantly, we assume that the meta-population exists in a world where $\mathbf{B_2}$ maintains its outsized capacity to confer punishment-costs on others and to resist their punishment efforts. When norm 2 is the only norm in the meta-population, then this outsized punishment capacity now provides the resident strategies with an advantage in conflict. From this perspective, Blue agents are now the "dominant" type in a colloquial sense. Correspondingly, a far less stringent criterion must hold for $\mathbf{B_2}$ to coerce cooperation from a mutant Red norm 1 playing agent:

Lemma 3. *(proof in Appendix D) In a meta-population defined by resident* $\boldsymbol{B_2}$ *and* $\boldsymbol{R_2}$ *strategies that punish norm* 1 *adherents, the best-response for a Red mutant that plays norm* 1 *following conflict is to adopt norm* 2 *in the second and all future rounds so long as the cost of being punished by the* $n-1$ *resident strategies in its group exceeds the potential private (net) cost-loss of adopting* *norm* 2*:*

$$\overbrace{(p^{R,1}_{B,2}+k^{B,2}_{R,1})(n-m)+(p^{R,1}_{R,2}+k^{R,2}_{R,1})(m-1)}^{\textit{\textbf{R}}_{\textit{1}}\textit{'s cost of conflict with } n-1 \textit{ residents}} \\ > \\ \underbrace{(c^2_R-\frac{b^2_R}{n})-(c^1_R-\frac{b^1_R}{n})}_{\textit{Private (net) cost-loss to } \textit{\textbf{R}}_{\textit{1}} \textit{ in switching to its not-preferred norm, } 2} \tag{1}$$

Lemma 3 as well as the more stringent 2 necessarily implies that a $\mathbf{R_{1\to 2}}$ (and norm 2) cannot invade the meta-population (proof in Section Appendix E.1). A $\mathbf{R_{1\to 2}}$ mutant engages in fitness-diminishing conflict for a single round before adopting the resident Red strategy's behavior, i.e. the mutant receives identical payoffs to the resident strategy from the second round forwards. However, Lemma 3 being true assumes that the cost of conflict exceeds any private cost-savings from contributing to norm 1 instead of norm 2. Correspondingly, the norm 1 playing $\mathbf{R_{1\to 2}}$ mutant's fitness is always less than the resident $\mathbf{R_2}$'s fitness. Because no norm change is possible, no benefits will enter into norm 2s stability criterion. However, that no norm change is possible is a function of the relative punishment parameters: A strategy that conforms to the dominant strategy's preferred behavior will always have higher fitness than a strategy that refuses to.

3.2.2. Stochastic Conflict Resolution

We can similarly examine the stability of norm 2 and $\mathbf{R_1}$ to norm 1 and $\mathbf{R_1}$ invasion when conflict is resolved by the stochastic mechanism, $\boldsymbol{\tau}$. We predict that now because normative change is possible, the the benefit of changing of normative change, to Red types, will enter the stability criterion, again, to the extent that $\boldsymbol{\tau}$ makes norm 1 rulings.

Proposition 5. *(proof in Section Appendix E.2) Norm* 2 *(and* $\boldsymbol{R_2}$*) is stable* *to a norm* 1 *(and* $\boldsymbol{R_1}$*) invasion so long as the probability of* $\boldsymbol{\tau}$ *making a norm* 1 *ruling times the change in benefits from adopting norm* 1 *to Red agents, plus any private-cost benefit to Red agents for adopting norm* 1 *is less than the cost to* $\boldsymbol{R_1}$ *of the first round of conflict plus any risk of ongoing future conflict:*

$$\begin{gathered}
[1-\omega\,(\tau_1+\tau_2)]\left[(p^{R,1}_{B,2}+k^{B,2}_{R,1})(n-m)-(p^{R,1}_{R,2}+k^{R,2}_{R,1})(m-1)\right] \\
> \\
\omega\tau_1(b^1_R-b^2_R)+(1-\omega\tau_2)(c^2_R-c^1_R)+[1-\omega\,(\tau_1+\tau_2)]\left(\frac{b^1_R-b^2_R}{n}\right)
\end{gathered} \tag{2}$$

This criterion is essentially identical to that of Proposition 4 studying the stability of norm 1. This is to be expected because the only factors impacting normative change when $\boldsymbol{\tau}$ exists are $\boldsymbol{\tau}$'s characteristics and these are broadly symmetrical. We assume that $\boldsymbol{\tau}$ does not interact differently with norm 1 or norm 2 bearing agents but rather rules in either set of agents' favor with some non-zero probability.

4. Discussion

Perhaps the key result of our model is that in permitting normative conflict and, importantly, for conflict to be resolved, something like a group-level benefit has entered into the stability criterion of a norm: Propositions 3 and 4. How should we interpret this result? The mechanistic model producing Proposition 4 perhaps provides the clearest results. Assuming a stable process of normative change, then *to the degree* that normative change is possible (τ_1, τ_2), benefits will enter into a norm's stability criterion. While in assuming conformity to $\boldsymbol{\tau}$'s rulings, we have ignored the endogenous nature of conflict resolution, that endogenous nature is self-evident. Our two models demonstrate that it is the various factors–technological, social, cultural, etc.–defining agents' relative success when seeking normative change that determine social evolution and who it benefits. Although our initial model characterized normative change as the outcome of something like political conflict–strikes, boycotts, or any factional implementation of costs–our stochastic model encourages a broader conception of normative change that also includes stabilized legal and political processes that may be biased, implicitly or explicitly, in favor of the interests of some societal actors over others (Pirie, 2013).

We point to the appearance of a contradiction to highlight the importance of coercion in normative change. When we enforce deterministic conflict resolution, the stability criterion of norm 1 depends on a group-level benefit (Section 3.1.1) yet the stability of norm 2 does not (Section 3.2.1)? The distinction stems from whether the invading norm alters the punishment capacity of an invading mutant or the resident strategies. In the former case, when the novel norm grants a mutant strategy an out-sized capacity to resist and inflict punishment costs, then, so long as the mutant strategy can coerce cooperation, then the resident norm will be unstable according to the criteria in Propositions 2 and 3 and benefits will enter into a norm's stability criterion. However, should the resident strategies' punishment capacity increase from the novel norm then normative change will be impossible. When normative change is impossible, we are squarely rooted in a world where benefits are completely irrelevant to a norm's stability criterion. In such a world, a mutant can only defect, but never coerce cooperation, and it has no reason to engage in normative conflict. One of the purposes of our stochastic model (Sections 3.1.2 and 3.2.2) was to demonstrate that so long as $\tau_1, \tau_2 > 0$, then normative change will always be possible, and the benefits of normative change will always enter into a norm's stability criterion.

We emphasize, however, that though benefits enter into a norm's stability criterion, in no sense do our results upend Boyd and Richerson's (1992) finding that punishment can stabilize norms irrespective of their benefits. Instead, we think our results expands on and qualifies previous findings by considering different contexts. First of all, we emphasize that no meta-population group-level benefit ever enters into a norm's stability criterion, but that only a type-specific benefit does. In fact, our formalism does not permit us to speak directly or immediately of group-level benefits and would require some more sophisticated social-welfare or cultural-fitness aggregating approach. Correspondingly, we strongly emphasize that no altruism between sub-populations or types can be assumed. Absent additional mechanisms between groups, such as overlapping social ties (Colson, 1953; Gluckman, 1955), costly-punishment will not drive culturally endogamous, yet interacting, types to be altruistic to one another. Additionally, it is not obvious when a mutant strategy should be committed to normative change and willing to engage in punishment to enforce novel norms. Our model assumes that a norm will interact with type to alter punishment capacity, but to some degree, this is simply a requisite of our formalism. In actuality, when agents participate in normative conflict will be a function of a society's specific culture and institutions. A less abstract model of normative change should almost certainly entail behavioral contexts and social dilemmas beyond punishment games that reflect these cultural and institutional characteristics. For instance, should societies consist of patron-client networks (Eisenstadt and Roniger, 1980; Mair, 1961; Scott, 1972), then the relationship between our two sub-populations will be characterized by specific norms, institutions, and corresponding psychologies (Henrich et al., 2010; Muthukrishna et al., 2021) that would need to be considered to understand the dynamics of normative change.

Our model also necessitated novel assumptions compared to previous models to permit our result. The key assumption is the presence of distinct ”preference” or fitness orderings defined by endogamous, but interacting, sub-populations.

We conceptualize this as a process of "social differentiation". While we do feel that previous models have failed to acknowledge the omnipresence of social differentiation in both cultural and genetic evolution, it is easy to argue that these previous models simply consider those cases where all individuals, regardless of their differences, share homogeneous preference orderings over norms. Our model then considers those cases where social differentiation produces distinct preference orderings and novel norms produce distinct punishment capacities. In such cases, particularly in a world without stable conflict resolution institutions, normative change can occur whenever a mutant adopt a novel norm is able to coerce cooperation with this norm. To reiterate, in order to propagate, the mutant norm-player must be able to impose larger costs on the resident strategy of its type than all group-members can impose on it. However, in the meta-population defined by norm 1, a single Blue mutant attempting to implement a novel norm will be immediately punished by $n - m - 1$ agents of its own type. This means that social differentiation in punishment capacity at the type-level is insufficient to permit normative change. Instead, the novel norm must interact with agents' type-specific punishment capacities *such that* the mutant can coerce cooperation. Future work should study whether more general formalisms permitting "social differentiation" exist and whether these alternative formalisms produce results distinct from our own.

Our deterministic model suggests that the various factors defining the distribution of conflict-parameters are the "prime movers" in normative change and necessitate more direct study than cultural evolutionists have allocated them (Nunn, 2021). In other words, when novel behaviors do not produce agents defined by heterogeneous punishment capacities, norms are stabilized by punishment arbitrarily of their benefits (Boyd and Richerson, 1992) and social change occurs on slower time-scales through cultural group selection mechanisms. However, should such a condition hold then normative conflict becomes the engine of social change at the intra-societal level. In a world defined by stabilized rules concerning normative change (Hadfield and Weingast, 2012; Hart, 1994), such social differentiation need not occur to permit normative change. If secondary rules about norm-change exist and individuals are incentivized to conform to the output of such rules when applied to specific cases (Hadfield and Weingast, 2012), the criterion of a norm's stability is still quite similar to the deterministic case. A norm is only stable if the cost of conflict over it, plus any future possible ongoing conflict, assuming the dispute resolver permits it, exceeds the probability of the dispute resolver ruling in favor of the novel norm multiplied by its relative benefit to the status-quo norm (Proposition 4). However, assuming this case merely postpones the question of conflict. The presence of effect of $\boldsymbol{\tau}$ immediately constructs incentives for agents to seek to influence and manipulate the organizations and institutions that produce legal rulings and normative classifications (Caldwell, 2014; Eder, 2005).

Similarly, both our deterministic and stochastic models suggest that norm-stability is always contingent on how agents' heterogeneous (1) interests or preferences interact with their heterogeneous (2) punishment capacities. In other words, the stability and presence of a norm indicates that *either* some subset of actors exists that is capable of enforcing an alternative norm, but does not perceive it to be in its shared interest to do so, or that no such subset of actors exists and the benefits of the norm cannot be assumed.

We think that it is important to emphasize that in the domain of normative and institutional change, the perception of interests is not a simple topic. In lab experiments, human participants do not spontaneously anticipate the equilibrium outcomes of extremely simple institutions and instead must experience those institutions directly to understand how they structure incentives (Gurerk et al., 2006). Historically and anecdotally, several of the founding parties to the United States Constitution inaccurately predicted its (in)stability (Rasmussen, 2021). It is not clear that several hundred years of experience with the U.S. constitution has improved our ability to predict or anticipate its impact on political dynamics either. Arguably, this will be the case for any institution or norm that either functions in a complex way or that has complex consequences or provides sparse reward (positive or negative) signals to agents. Agents engaging in normative conflict on the basis of anticipated rewards necessarily rely on some, arguably largely culturally evolved, model of the social, political, and economic world they are acting in and seek to change. In other words, shifting equilibria is an inherently imaginary and creative process and requires coalition architects and members seeking normative and institutional change to imagine both what is wrong with the current world as well as what new possible social world and equilibria (1) exist, (2) whether such equilibria can even reached, and (3) what path permits access to said equilibria (Wolcott, 2022). One wonders whether such stories are ever veridical and historical work indicates they can often be wrong (Scott, 1998). Correspondingly, the various factors that impact what stories individuals tell, believe, and spread about norms and institutions may hold a significant role in understanding both past, contemporary, and future normative and institutional change (DellaVigna and Kaplan, 2007; Field, 1976; Galletta and Ash, 2019; Hochschild, 2016; Luhrmann, 2020; Martin and Yurukoglu, 2017; Simonov et al., 2020).

We must also confront the question of group-selection. As previously mentioned, a mutant with an outsized ability to inflict and resist punishment costs may be able to impose its preferred norms. This is true regardless of how the norm impacts group-members of another type. Again, history points to punishment being used in two distinct capacities: both to upend inequalities, rents, and privileges as well as to impose and enforce them, somewhat analogous to Sections 3.1 and 3.2. One might quickly disregard political behavior and conflict as only able to produce self-interested outcomes and argue that it is only in light of group-selective forces that the set of agents able to enforce their preferred norms make prosocial concessions (Richerson et al., 2016). There is surely some degree of truth to this argument, if we assume even minimal capabilities of forward-looking behavior as well as awareness of what concessions are demanded.

However, the results of our model make several indications that the above story concerning group-selection must be studied more directly. First of all, in light of the fact that norms can be stabilized irrespective of their benefits (Boyd and Richerson, 1992; Gintis et al., 2001; Panchanathan and Boyd, 2004), social differentiation provides a path for agents to implement norm change that, depending on the broader institutions characterizing a society *may or may not* be prosocial. Secondly, normative conflict can be a path to equality and provides a pathway for types within a population to upend harmful inequalities; such inequality is hardly rare in ethnographic or historical work (Alfani, 2021; Bello, 2019; Flannery and Marcus, 2012; Hayden, 2016; North et al., 2009). More complicated models, characterized by more elaborate institutional characteristics calibrated to specific historical and ethnographic cases, would be needed to understand when normative change produces such prosocial shifts and when it steps into the construction of inequality.

With respect to the argument that cultural group selection will select for prosocial equilibria regardless of intra-societal forces, we emphasize that it is the intra-societal forces that we study that produce the equilibria group-selective forces both consume and interact with. While group-selection may interact with intra-societal forces to force the powerful to make prosocial concessions, it may also interact with intra-societal forces to limit the ability of the powerful to make said concessions. Intra-elite conflict for example, a major engine of history (Goldstone, 2016), could either incentivize subsets of elites to grant prosocial concessions in order to build novel coalitions permitting their victory in conflict with other elites *or* it could push elites to behave in less forward-looking ways, restricting their ability to offer prosocial concessions. Both of these hypotheses are reasonable and likely occur under different conditions. Such topics necessitate theoretical and empirical study by cultural evolutionists and characterize the very processes that upended chimpanzee-like dominance hierarchies and fundamentally altered the nature of sociality and social-selection in the human-lineage, ultimately resulting in the evolution of a fundamentally novel and normative organism (Boehm et al., 1993; Boehm, 1999; Boyd and Richerson, 1992; Boyd et al., 2010; Boyd, 2019; Panchanathan and Boyd, 2004). Because intra-societal forces generally act more rapidly and with greater strength than inter-group forces, failing to investigating how processes of normative conflict determine equilibria means that we cannot characterize the variation that group-selection consumes. This in turn implies that should such intra-societal forces also interact with group-selective ones, we cannot fully characterize the dynamics and outcomes of group-selection.

In light of our model and the above arguments, we feel that future research must explore (1) what determines the characteristics of normative conflict and its outcomes, i.e. when does it occur and when does it produce prosocial benefits rather than merely construct inequalities and privileges and (2) how do such normative phenomena interact with cultural group selection. The first of these questions can be rephrased as: Under what conditions do goal-seeking actors produce group-level benefits? Such a question for instance asks us to understand how it is that chimp like dominance hierarchies came to be inverted in human evolution by coalitions of the "weak" (Boehm et al., 1993; Boehm, 1999). Alternatively, what cognitive, social, technological, and institutional factors permit coalitions to be constructed, define which coalitions are likely to be constructed, and determine the outcomes of conflict among coalitions, ultimately determining what norms are stabilized. For instance, actors and organizations providing normative classifications (legal-order) may facilitate the emergence of normative coalitions (Hadfield and Weingast, 2012) as do factors impacting signaling outcomes (Boyd et al., 2010) and prestige transmission (Henrich et al., 2015). However, these processes are themselves sites of contention. Law itself is used opportunistically and strategically by social, political, and legal actors (Caldwell, 2014; Conley and O'Barr, 1990; Dixit, 2011; Eder, 2005; Lanni, 2016; Nader, 1997; Nader and Todd, 1978; Pirie, 2013). The second of our above questions considers how similar conflict processes within societies may determine the dynamics of group-selection, limiting or encouraging the spread of prosocial institutions. Actors and institutions actively construct their environments and the societies and agents populating them (for instance Spruyt, 1996), which would again determine the dynamics and outcomes of group-selection. For instance, restrictions on population movement by the powerful are an obvious impediment to individuals' ability to "exit" from groups with low-payoff equilibria to those with higher-payoff equilibria (Boyd and Richerson, 2009; Hirschman, 1970). While patents may benefit society broadly (Richerson et al., 2016) conditional on some broader set of institutions, can it be merely assumed that of the set of potential patent-rules, those that are socially optimal are implemented? A group-selection model that ignored those processes identified in our model would be forced to make such a case. Our model suggests that norms and institutions first must pass through a "sieve" where those preferred by agents with outsized punishment capacity are then subject to group selective forces and that such group selective forces may themselves be subject to the goals of powerful actors.

Finally, the implementation of cooperative and normative behaviors in artificial intelligence systems is a growing area of research (Dafoe et al., 2021). In light of cultural evolutionary models, recent attempts in multi-agent reinforcement learning have begun to make use of exogenously enforced punishment in order to stabilize costly-norms (Vinitsky et al., 2022; Yaman et al., 2022). Our results, because they expand on these very same cultural evolutionary models, speak to future versions of multi-agent reinforcement systems where punishment and conflict costs will inevitably be endogenized to permit more dynamics and general learning. Should agents learn roles and characteristics that impact their ability to engage in and resist punishment, our model will apply. Should our model apply, then which, if any, equilibria such systems settle on may reflect the interests of agents with outsized punishment capacities as opposed to a benefit-agnostic set of equilibria (Boyd and Richerson, 1992).

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.

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, 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.

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

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.

M. Derex, J. F. Bonnefon, R. Boyd, A. Mesoudi, Causal understanding is not necessary for the improvement of culturally evolving technology, Nature Human Behaviour 3 (2019) 446–452. doi:10.1038/s41562-019-0567-9.

R. Köster, D. Hadfield-Menell, R. Everett, L. Weidinger, G. K. Hadfield, J. Z. Leibo, Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents, Proceedings of the National Academy of Sciences 119 (2022) e2106028118. doi:10.1073/pnas.2106028118.

J. Barrett, Punishment and disagreement in the state of nature, Economics & Philosophy 36 (2020) 334–354. doi:10.1017/S0266267119000233.

F. Bray, The Rice Economies: Technology and Development in Asian Societies, Blackwell, New York, 1986.

G. Noblit, The origin and evolution of Chinese lineages (2021). doi:10.31235/osf.io/bq8ge.

D. Rustagi, S. Engel, M. Kosfeld, Conditional cooperation and costly monitoring explain success in forest commons management, Science 330 (2010) 961–965. doi:10.1126/science.1193649.

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.

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.

S. Ogilvie, The European Guilds: An Economic Analysis, Princeton University Press, Princeton, NJ, 2021.

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.

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.

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.

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

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.

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.), Foundations of Social Inequality, Fundamental Issues in Archaeology, Springer US, Boston, MA, 1995. doi:10.1007/978-1-4899-1289-3.

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.

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.

N. Buitron, H. Steinmüller, State legibility and mind legibility in the original political society, Religion and Society 12 (2021) 39–55. doi:10.3167/arrs.2021.120104.

C. Lévi-Strauss, Social and psychological aspect of chieftainship in a primitive tribe: The Nambikuara of northwestern Mato Grosso, Transactions of the New York Academy of Sciences, series II Vol. 7 (1945) 16–32.

R. H. Lowie, Some aspects of political organization among the American Aborigines, The Journal of the Royal Anthropological Institute of Great Britain and Ireland 78 (1948) 11. doi:10.2307/2844522. arXiv:2844522.

M. Mauss, Seasonal Variations of the Eskimo: A Study in Social Morphology, Psychology Press, 2004.

M. Sahlins, The original political society, HAU: Journal of Ethnographic Theory 7 (2017) 91–128. doi:10.14318/hau7.2.014.

D. Wengrow, D. Graeber, Farewell to the ‘childhood of man’: Ritual, seasonality, and the origins of inequality, Journal of the Royal Anthropological Institute 21 (2015) 597–619. doi:10.1111/1467-9655.12247.

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.

J. Knight, Institutions and Social Conflict, Cambridge University Press, Cambridge, U.K., 1992.

R. L. Kendal, N. J. Boogert, L. Rendell, K. N. Laland, M. Webster, P. L. Jones, Social learning strategies: Bridge-building between fields, Trends in Cognitive Sciences 22 (2018) 651–665. doi:10.1016/j.tics.2018.04.003.

D. Poulin-Dubois, P. Brosseau-Liard, The developmental origins of selective social learning, Current Directions in Psychological Science 25 (2016) 60–64. doi:10.1177/0963721415613962.

L. Rendell, L. Fogarty, W. J. Hoppitt, T. J. Morgan, M. M. Webster, K. N. Laland, Cognitive culture: Theoretical and empirical insights into social learning strategies, Trends in Cognitive Sciences 15 (2011) 68–76. doi:10.1016/j.tics.2010.12.002.

F. H. Knight, Risk, Uncertainty and Profit, Houghton Mifflin, Boston, 1921.

J. E. LeDoux, J. Moscarello, R. Sears, V. Campese, The birth, death and resurrection of avoidance: A reconceptualization of a troubled paradigm, Molecular Psychiatry 22 (2017) 24–36. doi:10.1038/mp.2016.166.

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

C. Heyes, Précis of cognitive gadgets: The cultural evolution of thinking, Behavioral and Brain Sciences 42 (2019). doi:10.1017/S0140525X18002145.

M. Muthukrishna, J. Henrich, E. Slingerland, Psychology as a historical science, Annual Review of Psychology 72 (2021) 717–749. doi:10.1146/annurev-psych-082820-111436.

L. J. Savage, The Foundations of Statistics, Dover Publications, New York, NY, 1954.

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.

M. Taylor, The Possibility of Cooperation, Cambridge University Press, Cambridge, UK, 1987.

J. Henrich, R. Boyd, Division of labor, economic specialization, and the evolution of social stratification, Current Anthropology 49 (2008) 715–724. doi:10.1086/587889.

T. Earle, M. Spriggs, Political economy in prehistory: A Marxist approach to Pacific sequences, Current Anthropology 56 (2015) 515–544. doi:10.1086/682284.

J. E. Arnold, S. Sunell, B. T. Nigra, K. J. Bishop, T. Jones, J. Bongers, Entrenched disbelief: Complex hunter-gatherers and the case for inclusive cultural evolutionary thinking, Journal of Archaeological Method and Theory 23 (2016) 448–499. doi:10.1007/s10816-015-9246-y.

H. L. A. Hart, The Concept of Law, second ed., Oxford University Press, New York, NY, 1994.

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.

M. Broom, J. Rychtář, Game-Theoretical Models in Biology, Chapman & Hall/CRC Mathematical and Computational Biology, Chapman & Hall/CRC, Boca Raton, FL, 2013.

P. D. Taylor, Evolutionarily stable strategies with two types of player, Journal of Applied Probability 16 (1979) 76–83. doi:10.2307/3213376.

S. Mathew, How the second-order free rider problem is solved in a small-scale society, American Economic Review 107 (2017) 578–581. doi:10.1257/aer.p20171090.

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.

F. Pirie, The Anthropology of Law, Oxford, New York, NY, 2013.

E. Colson, Social control and vengeance in Plateau Tonga Society, Africa: Journal of the International African Institute 23 (1953) 199–212. doi:10.2307/1156280. arXiv:1156280.

M. Gluckman, The Judicial Process among the Barotse of Northern Rhodesia, The Free Press, Glencoe, 1955.

S. N. Eisenstadt, L. Roniger, Patron-client relations as a model of structuring social exchange, Comparative Studies in Society and History 22 (1980) 42–77. doi:10.1017/S0010417500009154.

L. P. Mair, Clientship in East Africa, Cahiers d'Études Africaines 2 (1961) 315–325.

J. C. Scott, Patron-client politics and political change in Southeast Asia, American Political Science Review 66 (1972) 91–113. doi:10.2307/1959280.

J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world?, Behavioral and Brain Sciences 33 (2010) 61–83. doi:10.1017/S0140525X0999152X.

N. Nunn, History as evolution, in: A. Bisin, G. Federico (Eds.), The Handbook of Historical Economics, Elsevier, San Diego, CA, 2021, pp. 41–91. doi:10.1016/B978-0-12-815874-6.00010-1.

E. Caldwell, Social change and written law in early Chinese legal thought, Law and History Review 32 (2014) 1–30. doi:10.1017/S0738248013000606.

W. Eder, The political significance of the codification of law in archaic societies: An unconventional hypothesis, in: Social Struggles in Archaic Rome, Blackwell, Oxford, UK, 2005, pp. 239–267. doi:10.1002/9780470752753.ch10.

D. C. Rasmussen, Fears of a Setting Sun: The Disillusionment of America's Founders, Princeton University Press, Princeton, NJ, 2021.

V. W. Wolcott, Living in the Future: Utopianism and the Long Civil Rights Movement, University of Chicago Press, Chicago, IL, 2022.

J. C. Scott, Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed, Yale University Press, New Haven, 1998.

S. DellaVigna, E. Kaplan, The Fox News effect: Media bias and voting, The Quarterly Journal of Economics 122 (2007) 1187–1234. doi:10.1162/qjec.122.3.1187.

D. Field, Rebels in the Name of the Tsar, Houghton Mifflin, Boston, MA, 1976.

S. Galletta, E. Ash, How cable news reshaped local government, 2019.

A. R. Hochschild, Strangers in Their Own Land: Anger and Mourning on the American Right, The New Press, New York, NY, 2016.

T. M. Luhrmann, How God Becomes Real: Kindling the Presence of Invisible Others, Princeton University Press, Princeton, NJ, 2020.

G. J. Martin, A. Yurukoglu, Bias in cable news: Persuasion and polarization, American Economic Review 107 (2017) 2565–2599. doi:10.1257/aer.20160812.

A. Simonov, S. K. Sacher, J.-P. H. Dubé, S. Biswas, The persuasive effect of Fox News: Non-compliance with social distancing during the Covid-19 pandemic, 2020. doi:10.3386/w27237.

G. Alfani, Economic inequality in preindustrial times: Europe and beyond, Journal of Economic Literature 59 (2021) 3–44. doi:10.1257/jel.20191449.

J. A. Goldstone, Revolution and Rebellion in the Early Modern World: Population Change and State Breakdown in England, France, Turkey, and China, 1600-1850, 25th anniversary edition ed., Routledge,Taylor & Francis Group, an Informa business, New York, 2016.

J. M. Conley, W. M. O'Barr, Rules versus Relationships: The Ethnography of Legal Discourse, University of Chicago Press, Chicago, IL, 1990.

A. K. Dixit, Lawlessness and Economics: Alternative Modes of Governance, Princeton University Press, Princeton, NJ, 2011.

A. Lanni, Law and Order in Ancient Athens, Cambridge University Press, New York, 2016.

L. Nader (Ed.), Law in Culture and Society, University of California Press, Berkeley, 1997.

L. Nader, H. Todd, F. (Eds.), The Disputing Process: Law in Ten Societies, Columbia University Press, New York, 1978.

H. Spruyt, The Sovereign State and Its Competitors, Princeton University Press, Princeton, New Jersey, 1996.

A. O. Hirschman, Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States, Harvard University Press, Cambridge, 1970.

A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, T. Graepel, Cooperative AI: Machines must learn to find common ground, Nature 593 (2021) 33–36. doi:10.1038/d41586-021-01170-0.

E. Vinitsky, R. Köster, J. P. Agapiou, E. Duéñez-Guzmán, A. S. Vezhnevets, J. Z. Leibo, A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings, 2022. doi:10.48550/arXiv.2106.09012. arXiv:arXiv:2106.09012.

A. Yaman, J. Z. Leibo, G. Iacca, S. W. Lee, The emergence of division of labor through decentralized social sanctioning, 2022. arXiv:arXiv:2208.05568.

Appendix A. Proposition 1 Ongoing Punishment

The average fitness of the resident $\mathbf{B_1}$ is:

$$W(B_1) = w_B + \frac{b_B^1 - c_B^1}{1-\omega} \tag{A.1}$$

The fitness of the mutant $\mathbf{B_2}$ is:

$$W(B_2) = w_B + \frac{\frac{b_B^2}{n} + b_B^1\frac{n-1}{n} - c_B^2 - (p_{R,1}^{B,2} + k_{B,2}^{R,1})m - (p_{B,1}^{B,2} + k_{B,2}^{B,1})(n-m-1)}{1-\omega} \tag{A.2}$$

$W(B_2) > W(B_1)$ if and only if:

$$\begin{gathered}(c_B^1 - \frac{b_B^1}{n}) - (c_B^2 - \frac{b_B^2}{n}) \\ > \\ (p_{R,1}^{B,2} + k_{B,2}^{R,1})m + (p_{B,1}^{B,2} + k_{B,2}^{B,1})(n-m-1)\end{gathered} \tag{A.3}$$

Appendix B. Lemmas 1 and 2 Blue & Red Acquiescence

Assume again that $\mathbf{B_2}$ mutants sporadically arise in a population of resident strategies $\mathbf{B_1}$ and $\mathbf{R_1}$. After a single round of punishment between the two types, $\mathbf{B_1}$ can either acquiesce and adopt norm 2 and stop punishing and being punished by $\mathbf{B_2}$ or it can continue to play norm 1 and continue to engage in conflict with $\mathbf{B_2}$.

If the focal $\mathbf{B_1}$ agent adopts norm 2 it will avoid conflict with $\mathbf{B_1}$ but correspondingly will come to be punished by the resident strategies ($\mathbf{B_1}$ and $\mathbf{R_1}$), its fitness will be:

$$\begin{gathered} W(B_{1\to 2}) = w_B + \\ b_B^1 \frac{n-1}{n} + b_B^2 \frac{1}{n} - c_B^1 - (p_{B,2}^{B,1} + k_{B,1}^{B,2}) \\ \frac{\omega \left[b_B^1 \frac{n-2}{n} + 2\frac{b_B^2}{n} - c_B^2 - (p_{R,1}^{B,2} + k_{B,2}^{R,1})m - (p_{B,1}^{B,2} + k_{B,2}^{B,1})(n-m-2) \right]}{1-\omega} \end{gathered} \tag{B.1}$$

If the focal $\mathbf{B_1}$ continues to play 1, i.e. is the resident strategy, $\mathbf{B_1}$, it will continue to punish $\mathbf{B_2}$ but avoid punishment by the resident strategies in its group. Its fitness will be:

$$\begin{gathered} W(B_{1\to 1}) = w_B + \\ b_B^1 \frac{n-1}{n} + b_B^2 \frac{1}{n} - c_B^1 - (p_{B,2}^{B,1} + k_{B,1}^{B,2}) \\ \frac{\omega \left[b_B^1 \frac{n-1}{n} + b_B^2 \frac{1}{n} - c_B^1 - (p_{B,2}^{B,1} + k_{B,1}^{B,2}) \right]}{1-\omega} \end{gathered} \tag{B.2}$$

$W(B_1 \to 2) > W(B_1 \to 1)$ if and only if:

$$\begin{gathered} (p_{B,2}^{B,1} + k_{B,1}^{B,2}) + (c_B^1 - \frac{b_B^1}{n}) - (c_B^2 - \frac{b_B^2}{n}) \\ > \\ (p_{R,1}^{B,2} + k_{B,2}^{R,1})m + (p_{B,1}^{B,2} + k_{B,2}^{B,1})(n-m-2) \end{gathered} \tag{B.3}$$

As Boyd and Richerson (Boyd and Richerson, 1992) discuss in their appendix, these fitness functions are heuristics and actually depend on the full distribution of how agents punish and respond to punishment. Payoffs and selective-processes are not the sole determinants of this distribution and ”non-adaptive processes like mutation and nonheritable environmental variation” will also inform this distribution (Boyd and Richerson, 1992, p. 189). Some may cease punishing after two rounds of conflict, three, etc. Note also that we do not consider a classical game theoretic model whereby agents may understand, or hold beliefs, that they face similar incentives in which case the above criterion becomes easier to satisfy because multiple agents would simultaneously acquiesce to punishment.

An identical process occurs for the Red agents. Acquiescing resident Reds, labeled $\mathbf{R_{1\to2}}$, initially play norm 1 before switching to norm 2 once punished by $\mathbf{B_2}$s. $\mathbf{R_{1\to2}}$ is fitter than $\mathbf{R_{1\to1}}$ (the resident strategy) if and only if:

$$\begin{gathered}(p_{B,2}^{R,1} + k_{R,1}^{B,2}) + (c_R^1 - \frac{b_R^1}{n}) - (c_R^2 - \frac{b_R^2}{n}) \\ > \\ (p_{R,1}^{R,2} + k_{R,2}^{R,1})(m-1) + (p_{B,1}^{R,2} + k_{R,2}^{B,1})(n-m-1)\end{gathered} \tag{B.4}$$

Appendix C. Norm 1 Stability to $\mathbf{B_2}$ Invasion

Appendix C.1. Deterministic Conflict Resolution

An invasion criterion indicates when a strategy that is initially rare can propagate relative to resident strategies. Formally, we ask when the fitness of a $\mathbf{B_2}$ mutant is larger than the average fitness of the resident strategies. Because criterion 1 holds, we assume the resident strategy is $\mathbf{B_{1\to2}}$.

The fitness of the mutant $\mathbf{B_2}$ is given by:

$$\begin{gathered}W(B_2) = w_B + \\ b_B^1\frac{n-1}{n} + b_B^2\frac{1}{n} - c_B^2 - (p_{B,1}^{B,2} + k_{B,2}^{B,1})(n-m-1) - (p_{R,1}^{B,2} + k_{B,2}^{R,1})m + \\ \frac{\omega}{1-\omega}[b_B^1\frac{m}{n} + b_B^1\frac{n-m}{n} - c_B^2 - (p_{R,1}^{B,2} + k_{B,2}^{R,1})m]\end{gathered} \tag{C.1}$$

The second line gives the mutant's payoff from the first round of being the sole contributor to the second public good and engaging in conflict with $n-m-1$ Blue agents and m Red agents. The third line gives all future payoffs once the Blue agents have switched to the second public good because we assume that 1 holds. The $-(p^{B,2}_{R,1} + k^{R,1}_{B,2})m$ term indicates that conflict with the Red resident strategies continues because 2 does not hold..

The average $\mathbf{B_{1\to 2}}$ however will not encounter a mutant but rather only other $\mathbf{B_{1\to 2}}$s and $\mathbf{R_1}$s, both of which punish (and enforce) behavior 1. The average $\mathbf{B_{1\to 2}}$ payoff then is given by:

$$W(B_{1\to 2}) = w_B + \frac{b^1_B - c^1_B}{1-\omega} \tag{C.2}$$

Norm 1 is stable so long as:

$$\begin{aligned} (p^{B,2}_{B,1} + k^{B,1}_{B,2})(n-m-s) + \frac{(p^{B,2}_{R,1} + k^{R,1}_{B,2})m}{1-\omega} \\ < \\ \frac{s}{n}(b^2_B - b^1_B) + \frac{(c^1_B - c^2_B) + \omega\left[(b^2_B - b^2_B \frac{m}{n}) - (b^1_B - b^1_B \frac{m}{n})\right]}{1-\omega} \end{aligned} \tag{C.3}$$

Appendix C.2. Stochastic Conflict Resolution

Assume that if two parties engage in simultaneous punishment then conflict is stochastically resolved in the first round following conflict by a mechanism, $\boldsymbol{\tau}$. This mechanism can make three rulings. With probablity τ_1, norm $\tilde{2}1$ agents should adopt norm 1; with probability τ_2, norm 1 playing agents should adopt norm 2; and with probability $1 - \tau_1 - \tau_2$ conflict continues.

Our intention is not to fully model the process of conflict resolution, which we assume to be resolved cooperatively. Instead, our goal is to show the effect of conformity to a mechanism that can produce these three possible rulings on norm stability.

The fitness of the $\mathbf{B_2}$ mutant is given by:

$$\begin{gathered}
W(B_2) = w_B + b_B^1 \frac{n-1}{n} + b_B^2 \frac{1}{n} - c_B^2 - \\
(p_{B,1}^{B,2} + k_{B,2}^{B,1})(n-m-1) + (p_{R,1}^{B,2} + k_{B,2}^{R,1})m + \\
\frac{\omega}{1-\omega} \Bigg[\tau_1 (b_B^1 - c_B^1) + \tau_2 (b_B^2 - c_B^2) + \\
(1 - \tau_1 - \tau_2)(b_B^1 \frac{n-1}{n} + b_B^2 \frac{1}{n} - c_B^2 - \\
(p_{B,1}^{B,2} + k_{B,2}^{B,1})(n-m-1) + (p_{R,1}^{B,2} + k_{B,2}^{R,1})m) \Bigg]
\end{gathered} \tag{C.4}$$

The mutant's fitness is compared to the average Blue strategy, $\mathbf{B_1}$, who receives a norm 1 payoff for the full experience of its life and never encounters the mutant:

$$W(B_1) = w_B + \frac{b_B^1 - c_B^1}{1-\omega} \tag{C.5}$$

Norm 1 is stable in the population so long as $\mathbf{B_2}$ cannot invade, i.e. its fitness is lower than the resident strategy's fitness:

$$W(B_1) > W(B_2) \tag{C.6}$$

This is true so long as:

$$\begin{gathered}
\omega \tau_2 (b_B^2 - b_B^1) + (1 - \omega \tau_1)(c_B^1 - c_B^2) + [1 - \omega(\tau_1 + \tau_2)] \left(\frac{b_B^2 - b_B^1}{n} \right) \\
< \\
[1 - \omega(\tau_1 + \tau_2)] \left[(p_{B,1}^{B,2} + k_{B,2}^{B,1})(n-m-1) + (p_{R,1}^{B,2} + k_{B,2}^{R,1})m \right]
\end{gathered} \tag{C.7}$$

Appendix D. Lemma 3 Red Acquiescence Norm 2 Population

When norm 2 is the dominant norm, i.e. fixed in the meta-population, a mutant $\mathbf{R_1}$ agent can adopt two paths of play. It can either respond to punishment by the $n-1$ norm 2 strategies in its group with cooperation, adopting norm 2 on all future rounds, or it can continue to play norm 1 and continue to engage in conflict with the $n-1$ norm 2 group-members for all future rounds. In line with our previous notation, we label the former strategy $\mathbf{R_{1\to2}}$ and the latter $\mathbf{R_{1\to1}}$.

The fitness of $\mathbf{R_{1\to1}}$ is given by:

$$\begin{gathered}W(R_{1\to1}) = w_R + \\ \frac{b_R^2\frac{n-1}{n} + b_R^1\frac{1}{n} - c_R^1 - (p_{B,2}^{R,1} + k_{R,1}^{B,2})(n-m) - (p_{R,2}^{R,1} + k_{R,1}^{R,2})(m-1)}{1-\omega}\end{gathered} \tag{D.1}$$

The fitness of $\mathbf{R_{1\to2}}$ is given by:

$$\begin{gathered}W(R_{1\to2}) = w_R + \\ b_R^2\frac{n-1}{n} + b_R^1\frac{1}{n} - c_R^1 - (p_{B,2}^{R,1} + k_{R,1}^{B,2})(n-m) - (p_{R,2}^{R,1} + k_{R,1}^{R,2})(m-1) + \\ \frac{\omega\left[b_R^2 - c_R^2\right]}{1-\omega}\end{gathered} \tag{D.2}$$

The mutant $\mathbf{R_1}$ agent will be coerced to adopt norm 2 if and only if $W(R_{1\to2}) > W(R_{1\to1})$:

$$\begin{gathered}(p_{B,2}^{R,1} + k_{R,1}^{B,2})(n-m) + (p_{R,2}^{R,1} + k_{R,1}^{R,2})(m-1) \\ > \\ (c_R^2 - \frac{b_R^2}{n}) - (c_R^1 - \frac{b_R^1}{n})\end{gathered} \tag{D.3}$$

This criterion is distinct from that found in Lemma 2 because Lemma 2 is defined for when the resident strategies are norm 1 agents whereas now, norm 2 playing agents are the only strategies found in the population. Correspondingly, a mutant $\mathbf{R_{1\to2}}$ faces no punishment once it switches in the second round unlike when $\mathbf{R_{1\to2}}$ occupied an ecology defined by other punishing $\mathbf{R_{1\to1}}$ agents as well.

Appendix E. Norm 2 Stability to $\mathbf{R_{1\to 2}}$ Mutant

Appendix E.1. Deterministic Conflict Resolution

If Lemma 3 holds then a mutant Red agent that initially plays norm 1 will adopt norm 2 after it is punished in the first round by the resident norm 1 playing strategies. Call this mutant $\mathbf{R_{1\to 2}}$. Note, as discussed in the main text, such an agent will never be able to coerce its group-members to adopt norm 1 and we should not expect the following stability criterion to consider group- (or type-) level benefits.

In order to assess norm 2's stability, we must consider when a mutant $\mathbf{R_{1\to 2}}$ can invade a resident $\mathbf{R_2}$ population. As discussed in the main text, norm 1's invasion depends on dynamics of the Red population, the population of agents that prefers public good 1, i.e. the type for which $b_R^1 > b_R^2$. We must ask then when the resident Red strategy that conforms to norm 2's fitness exceeds that of the mutant Red agent's.

The resident strategy's average fitness is given by:

$$W(R_2) = w_R + \frac{b_R^2 - c_R^2}{1-\omega} \tag{E.1}$$

In most groups no mutant exists and thus the average fitness is simply the full benefit of the public good 2 minus the cost of provisioning that good. The fitness of the mutant is given by:

$$\begin{gathered} W(R_{1\to 2}) = w_R + \\ b_R^2 \frac{n-1}{n} + \frac{b_R^1}{n} - c_R^1 - (p_{B,2}^{R,1} + k_{R,1}^{B,2})(n-m) - (p_{R,2}^{R,1} + k_{R,1}^{R,2})(m-1) + \\ \frac{\omega \left[b_R^2 - c_R^2 \right]}{1-\omega} \end{gathered} \tag{E.2}$$

The fitness of the resident strategy exceeds that of the mutant, $W(R_2) > W(R_{1\to 2})$ if and only if:

$$(p^{R,1}_{B,2}+k^{B,2}_{R,1})(n-m)+(p^{R,1}_{R,2}+k^{R,2}_{R,1})(m-1) \\ > \\ (c^2_R-\frac{b^2_R}{n})-(c^1_R-\frac{b^1_R}{n}) \tag{E.3}$$

Note that this an identical criterion to that defining Lemma 3. It states that the cost of being punished my the resident strategies must exceed the net-cost of switching from norm 1 to 2. However, the right hand side, $(c^2_R-\frac{b^2_R}{n})-(c^1_R-\frac{b^1_R}{n})$, may in fact be negative.

Appendix E.2. Stochastic Conflict Resolution

We now evaluate the stability of norm 2 to a $\mathbf{R_1}$ mutant when conflict is resolved by a mechanism, $\boldsymbol{\tau}$. The fitness of the resident Red strategy is identical to that in Section Appendix E.1. The mutant Red strategy's fitness is now given by:

$$\begin{gathered} W(R_1)=w_R+ \\ b^2_R\frac{n-1}{n}+\frac{b^1_R}{n}-c^1_R-(p^{R,1}_{B,2}+k^{B,2}_{R,1})(n-m)-(p^{R,1}_{R,2}+k^{R,2}_{R,1})(m-1)+ \\ \frac{\omega}{1-\omega}\Bigg[\tau_1(b^1_R-c^1_R)+\tau_2(b^2_R-c^2_R)+ \\ (1-\tau_1-\tau_2)\Bigg[b^2_R\frac{n-1}{n}+\frac{b^1_R}{n}-c^1_R- \\ (p^{R,1}_{B,2}+k^{B,2}_{R,1})(n-m)-(p^{R,1}_{R,2}+k^{R,2}_{R,1})(m-1)\Bigg]\Bigg] \end{gathered} \tag{E.4}$$

Mirroring norm 1's stability criterion found in 4, norm 2 is stable so long as:

$$\begin{gathered} [1-\omega\,(\tau_1+\tau_2)]\left[(p^{R,1}_{B,2}+k^{B,2}_{R,1})(n-m)-(p^{R,1}_{R,2}+k^{R,2}_{R,1})(m-1)\right] \\ > \\ \omega\tau_1(b^1_R-b^2_R)+(1-\omega\tau_2)(c^2_R-c^1_R)+[1-\omega\,(\tau_1+\tau_2)]\left(\frac{b^1_R-b^2_R}{n}\right) \end{gathered} \tag{E.5}$$