Mankat, Fabian

Conference Paper

Cooperation, Norms, and Gene-culture Coevolution

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage"

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Mankat, Fabian (2023) : Cooperation, Norms, and Gene-culture Coevolution, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage", ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/277666

Cooperation, Norms, and Gene-culture Coevolution

Fabian Mankat*

February 28, 2023

Abstract

This paper investigates how human societies sustain positive levels of cooperation through the transmission and enforcement of norms. To do so, it introduces an evolutionary model that distinguishes between three distinct dynamic dimensions: behavior, norms, and approval preferences. These dimensions differ concerning their speed and nature of evolution. Whereas behavior evolves at the individual level through utility enhancement, norms evolve at the cultural level through peer interactions and socialization. Preferences are (at least partly) biologically inherited and therefore transmitted from parents to their offspring. The model suggests that if cultural and biological reproductive fitnesses are determined not only by material factors but also by social ones, then an interplay of social disapproval mechanisms can explain the persistence of norm-driven cooperation and heterogeneity regarding cooperative behavior and attitudes across situations and individuals.

*University of Kassel, Faculty of Economics and Management

Contents

1 Introduction **3**

2 Theoretical Framework **6**

3 Behavior **10**

3.1 Evolutionary Framework 10

3.2 Equilibrium Analysis 11

4 Norms **14**

4.1 Evolutionary Framework 14

4.2 Equilibrium Analysis 15

4.2.1 Homogeneous Approval Preferences 16

4.2.2 Heterogeneous Approval Preferences 19

4.3 Discussion 24

5 Approval Preferences **26**

5.1 Evolutionary Framework 26

5.2 Equilibrium Analysis 27

5.2.1 The Potential Biological Equilibrium 28

5.2.2 Stability-Inducing Equilibrium Culture 31

5.2.3 The Biological Equilibrium and Prevailing Social Norms 35

5.3 Discussion 43

6 Concluding Remarks **45**

A Supplementary Analysis **I**

A.1 Behavioral Equilibria I

A.2 Additional Cultural Equilibria I

A.3 Decreasing Costs of Contribution V

A.4 The Weight of Social Approval on Reproductive Fitnesses VII

A.5 Notes on Social Disapproval for Non-Conformity VIII

B Proofs and Additional Formal Results **X**

B.1 Behavior X

B.2 Norms XV

B.3 Preferences XXI

1 Introduction

Human societies uphold cooperation among non-related individuals, even if such behavior is relatively costly to the individuals themselves. Bridging the divergence of self-interest and cooperation is often accredited to the existence and transmission of informal institutions such as social norms (e.g., Elster 1989; Ostrom 2000). A social norm captures a society's shared understanding of what behavior is appropriate in a particular situation (Crawford and Ostrom 1995). Individuals follow social norms due to the threat of social sanctions such as disapproval by others (Fehr and Fischbacher 2004; Voss 2001). Moreover, individuals often go out of their way to act according to what they consider morally right. Such self-based standards of behavior are often referred to as personal norms (Nyborg 2018). They guide an individual's behavior through inner feelings such as guilt and self-perception (Thøgersen 2006). Acknowledging the existence of norms and their impact on individuals' decision-making can explain cooperative behavior in different situations.[1] However, it raises new questions regarding their underlying evolutionary foundations. In particular, how can a society ensure the transmission and enforcement of cooperation prescribing norms?

This paper contributes to answering this question. The main contribution to the existing literature is two-fold. First, to the best of my knowledge, the paper present the first model that studies the co-evolution of personal norms, social norms, and approval preferences. Thereby, it endogenizes the formation of norms and the mechanism that enforces them while accounting for the rich set of dynamic inter-dependencies. Second, the analysis highlights the potential role different social disapproval mechanisms may play in shaping society's culture. We establish the interplay of social disapproval mechanisms as an explanation for the persistence of norm-driven cooperation and heterogeneity in cooperative behavior and attitudes across individuals and situations.

The evolutionary model of this paper consists of three distinct dynamic dimensions: behavior, norms, and approval preferences. Behavior evolves at the individual level through utility enhancements. Norms evolve at the cultural level through peer interactions and

[1] See among others Akerlof and Kranton (2005), Andreoni and Bernheim (2009), Bénabou and Tirole (2006, 2011), Bernheim (1994), Brekke et al. (2003), d'Adda et al. (2020), Figueres et al. (2013), Nyborg (2000), Nyborg and Rege (2003a), Rabin (1995), and Traxler (2010).

socialization institutions (Joseph Henrich and Gil-White 2001). Approval preferences are (at least partly) transmitted through biological reproduction (Chudek and Joseph Henrich 2011). The evolution of norms and preferences is driven by social status, a combination of material payoff and social approval. The underlying notion is that socially successful individuals have a greater impact on the opinion formation of their peers (Bowles and Gintis 1998) and are more likely to find mating partners (Buss and Schmitt 1993; Turke 1989). Thus, their personal norms and approval preferences spread in society. The distribution of personal norms specifies what the individuals generally regard as appropriate and, thus, defines the social norm (Carbonara et al. 2008; Cooter 1998). Social disapproval arises from three sources: social norm violation by acting against what society generally considers appropriate, personal norm non-conformity by holding conflicting moral views to others, and hypocrisy by engaging in behavior that conflicts with one's own personal norm.

The proposed framework gives rise to an evolutionary stable distribution of preferences, norms, and behavior, where norms and behavior vary across individuals and situations. Although approval preferences are possibly heterogeneous, in equilibrium society behaves as if it was homogeneous. Social disapproval for social norm violation provides individuals with incentives to cooperate at the behavioral level, favors norm evolution at the cultural level, and allows for norm-sensitive preferences at the biological level. It suffices to stabilize norm-driven cooperation if either norms or preferences are exogenous. Social disapproval for non-conformity stabilizes perfect social norms at the cultural level. Social disapproval for hypocrisy is responsible for the persistence of heterogeneous norms. It provides individuals with an additional cooperation incentive at the behavioral level. Moreover, it introduces an evolutionary advantage of preferences for self-approval if individuals cannot foresee the whole extent of their actions in terms of social disapproval. However, it negatively impacts the cultural fitness of individuals with cooperation prescribing personal norms that defect, which can hinder the preservation of norms if cooperation is very costly. The results shed some light on the role of heterogeneous environments. In particular, complete cooperation in some situations can explain the persistence of incomplete cooperation in other situations, and incomplete cooperation in some situations can secure large levels of cooperation in very costly situations.

The model closely relates to the evolutionary literature on norms that employs the *indirect evolutionary approach* as proposed by Güth and Yaari (1992). The underlying idea is that utility governs behavior, which determines the reproductive fitness of cultural and biological traits that, in turn, shape the utility function.[2] Mengel (2008) uses this method to analyze the cultural transmission of cooperation norms in non-integrated societies. Individuals are recurrently matched to interact in the prisoners' dilemma. Any individual who has internalized the cooperation norm experiences internal sanctions for defecting. Incomplete integration is modeled through a biased matching structure that favors alike individuals (in terms of norm internalization) to interact. Mengel (2008) finds that cooperation norms of intermediate strength can survive for high levels of integration and low institutional pressure. In contrast, strict norms require either low levels of integration or high institutional pressure. Alger and Weibull (2013) and Alger and Weibull (2016) also study evolutionary models that incorporate assortative matching. Rather than norm internalization, they focus on the evolution of preferences for complying with a certain moral norm. The moral norm is endogenous to the game and determined by what can be viewed as an application of Kant's categorical imperative: "what would maximize welfare given that everyone acted accordingly?". Similar to Mengel (2008), they find that assortative matching gives rise to the stability of norm-sensitive preferences. This paper complements these results by looking at situations where the material payoff depends on the behavior of the whole society rather than in-group peers. In such situations, assortative matching provides no evolutionary advantage to cooperative individuals, so additional explanations are needed.

Traxler and Spichtig (2011) present an evolutionary model that studies such a public goods game played at the societal level. The model endogenizes the dis-utility from sanctions

[2]Many other strands of literature look at norms in an evolutionary context. Azar (2004), Binmore and Samuelson (1994), Lindbeck et al. (1999), Nyborg and Rege (2003b), Rege (2004), Sethi and Somanathan (1996), and Young (1993, 1996, 2015) focus on the evolution of behavior to rationalize norm-compliance. Bezin (2019), Bisin, Topa, et al. (2004), Bisin and Verdier (2001), and Tabellini (2008) study cultural evolution through *rational socialization*, where parents rationally choose what values to transmit to their offspring. Panebianco (2016) introduces an evolutionary model of norms that incorporates persuasion of peers. Bowles and Gintis (1998), Robert Boyd and Peter J Richerson (2005, 1990), Joseph Henrich (2004), and Mitteldorf and Wilson (2000) propose group selection arguments where norms persist since they are group-advantageous. Beyond norms, this paper contributes to the general literature on the evolution of cooperation-inducing traits using the indirect evolutionary approach (see Bester and Güth (1998), Guttman (2003, 2013), Müller and Wangenheim (2019), and A. Poulsen and O. Poulsen (2006)among others).

for social norm violation. Moreover, the sanctions positively depend on society's overall level of contribution. This set-up gives rise to multiple behavioral equilibria. For every interaction in the public goods game, society reaches each behavioral equilibrium with some positive probability. Moreover, Traxler and Spichtig (2011) assume that the reproductive fitness of preferences is co-determined by material payoff and social sanctions. They find that evolution favors an intermediate degree of social sanction sensitivity since it allows individuals to behave flexibly and adapt behavior to the given environment. The analysis connects to Traxler and Spichtig (2011), in that it provides further insights on the role of heterogeneous environments. However, whereas Traxler and Spichtig (2011) introduce heterogeneity within a situation through different behavioral equilibria, this paper incorporates heterogeneous environments through the interactions across different situations. This paper also relates to Traxler and Spichtig (2011) in a structural way, since reproductive fitness of cultural and biological traits is co-determined by material factors and social ones. Moreover, it endogenizes the strength of social sanctions for social norm violation. However, rather than behavior, it focuses on the role of moral perceptions as a determinant of endogenous norm strength. In this respect, our works are complementary.

The remainder of this paper is as follows. Section 2 presents the the static theoretical framework. Section 3 and 4 analyze the evolution of behavior and norms in a single situation. Thereafter, section 5 endogenizes the evolution of preferences in multi-situational environments. Finally, section 6 discusses the results and provides an outlook for future research.

2 Theoretical Framework

The model consists of a large society of individuals $i \in \mathcal{I}$, who recurrently interact in different situations $\omega \in \Omega$. Each situation ω constitutes a public goods game, where individual i executes action $a_i^\omega \in \mathcal{A} = \{0, 1\}$. She either contributes to the public good ($a_i^\omega = 1$, 'cooperate') or not ($a_i^\omega = 0$, 'defect'). We denote the share of individuals that contribute in situation ω by ψ^ω. At times, we refer to it as the level of cooperation.

An individual i's approval preferences are indicated by her preference type $\theta_i = (\theta_{s_i}, \theta_{p_i}) \in$

$\Theta \subset \mathbb{R}^2_{\geq 0}$, where Θ is an arbitrarily large but finite set.[3] The vector $\lambda \in [0,1]^{|\Theta|}$ describes the distribution of approval preferences in society, where $\lambda_\theta \in [0,1]$ corresponds to the share of individuals i for whom $\theta_i = \theta$. The support of any distribution λ indicates which preference types exist in this population. Formally, $\text{supp}(\lambda) := \{\theta \in \Theta : \lambda_\theta > 0\}$. For obvious reasons, we require that $\sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta = 1$.

The personal norm $n_i^\omega \in \{0,1\}$ captures the behavior that i considers morally appropriate in situation ω. An individual i holds the *cooperation norm* for situation ω, $n_i^\omega = 1$, if she considers cooperation to be the only morally right thing to do in that situation. If individual i does not hold the cooperation norm, $n_i^\omega = 0$, then she considers all possible actions appropriate.[4] We will sometimes refer to the sub-population of individuals who hold the cooperation norm as *norm holders*. Analogously, we say that an individual is a *norm non-holder* if she does not consider cooperation the only morally right thing to do. Individuals communicate their personal norms to peers. We assume that this communication occurs truthfully (possibly due to a positive probability of being detected as a liar, which might lead to substantial social and material costs).[5]

A social norm for situation ω captures societies' shared understanding of appropriate behavior in that situation. In line with Cooter (1998) and Carbonara et al. (2008), a social norm is, thus, defined by the distribution of personal perceptions of morally acceptable behavior, namely the distribution of personal norms. ϕ^ω is the share of individuals that hold the cooperation norm for situation ω. If ϕ^ω is large, many individuals believe that cooperating is the only morally right thing to do, and we say it is a strong social norm. The vector $\phi \in [0,1]^{|\Omega|}$ captures a social norm ϕ^ω for each situation $\omega \in \Omega$.

Individuals participate in information sharing (gossip) about their peers' behavior and personal norms. We indicate the degree of gossip in society by $\delta \in \mathbb{R}_{\geq 0}$. We assume that

[3] In principle the set of preference types coincides with all possible elements of $\mathbb{R}^2_{\geq 0}$. However, this creates problems regarding the traceability of our evolutionary model, for which reason we need to adopt this simplifying assumption. Note that since we allow for Θ to be arbitrarily large, we do not impose any restrictions on which preference types exist or may occur.

[4] Since the analysis explicitly focuses on the evolution of pro-social norms that induce cooperation, we disregard norms that prescribe defection.

[5] Abeler et al. (2019) show in a meta-analysis that untruthful reporting occurs surprisingly little even if it is beneficial to an individual. Bašić and Quercia (2022) provide some evidence that truth-telling seems to be motivated by social concerns.

gossip δ is exogenous.[6]

Individual i's material payoff m_i^ω from situation ω is determined by her own action a_i^ω and the cooperation level ψ^ω according to the payoff-function $m^\omega : \mathcal{A} \times [0,1] \to \mathbb{R}$. Throughout, m^ω is assumed to be continuous and differentiable in its second argument.[7] To capture the nature of public goods games, the material payoff of an individual i increases in the share of others who contribute. Moreover, contributing to the public good is relatively costly. For simplification purposes, we assume that contributing in any situation ω becomes relatively more costly in the share of others who contribute.[8] The main results do not change if we relax this assumption (see appendix A.3).

Definition 2.1. [Material Payoff] $m_i^\omega(a_i^\omega, \psi^\omega) = m^\omega(a_i^\omega, \psi^\omega)$ where $\forall \psi^\omega \in [0,1]$:

1. $\frac{\partial m^\omega}{\partial \psi^\omega} > 0$,
2. $\Delta m^\omega(\psi^\omega) := m^\omega(1, \psi^\omega) - m^\omega(0, \psi^\omega) < 0$, and
3. $-\frac{\mathrm{d}\Delta m^\omega(\psi^\omega)}{\mathrm{d}\psi^\omega} < 0$.

Self-approval captures how an individual i evaluates her behavior based on her personal norms. An individual who holds the cooperation norm for ω and does not act accordingly experiences inner emotions such as guilt and loss of self-esteem.

Definition 2.2 (Self-approval)**.** $p_i^\omega(a_i^\omega, n_i^\omega) = (a_i^\omega - 1)n_i^\omega$.

Social approval captures how i is perceived by her peers and derives from three separate components. First, individuals are subject to social disapproval from social norm violation $(a_i^\omega - 1)v(\phi^\omega)$. This social disapproval arises as a consequence of acting inappropriate in the eyes of the public. It requires the social norm to be present and increases in it's strength such that $v(0) = 0$ and $\frac{\mathrm{d}v(\phi^\omega)}{\mathrm{d}\phi^\omega} > 0$. Second, individuals are subject to social disapproval from non-conformity $k(|n_i^\omega - \phi^\omega|)$. The more individuals in society hold moral views conflicting with

[6] In section 6 we briefly argue that gossip itself can be seen as a public goods game to which the results of this paper apply.

[7] By understanding material payoff in expected terms, the described framework can also capture public dilemma games played in smaller randomly matched groups (e.g., prisoner's dilemma), where ψ^ω is the expected action of a randomly chosen individual.

[8] Larger levels of contribution may either require more effort to contribute or decrease marginal benefits from the public good.

those of i, the greater i's disapproval from non-conformity. Thus, social disapproval from non-conformity increases the distance between i's personal norm and the social norm. If i's personal norm coincides with the social one, she experiences no such disapproval. Formally, $\frac{\partial k(|n_i^\omega - \phi^\omega|)}{\partial |n_i^\omega - \phi^\omega|} > 0$ and $k(0) = 0$. Third, individuals are subject to social disapproval from hypocrisy $(a_i^\omega - 1)n_i^\omega h$, where $h > 0$. An individual i is perceived as a hypocrite if she does not contribute to the public good despite holding the cooperation norm.

Definition 2.3 (Social Approval)**.** $s_i^\omega(a_i^\omega, n_i^\omega, \phi^\omega) = (a_i^\omega - 1)[v(\phi^\omega) + n_i^\omega h] + k(|n_i^\omega - \phi^\omega|)$.

Following Nyborg and Rege (2003b), we assume that peers partly express social disapproval towards i as a direct reaction to her behavior and personal norms. This occurs through gestures such as raised eyebrows or similar, which do not automatically imply substantial costs for the individuals expressing them (Rege 2004). Neither are they necessarily subject to a deliberate and/or conscious decision (Blau 1964; Gächter and Fehr 1999). Observing these gestures of disapproval affects an individual's well-being through negative emotions of feeling rejected and condemned.[9]

The expression of social disapproval for social norm violation occurs as a reaction to observing behavior. Similarly, expressing social disapproval for non-conformity occurs directly from observing a personal norm. We write social disapproval from social norm violation and non-conformity that is directly communicated to an individual as $\tilde{v}(\phi^\omega)$ and $\tilde{k}(|n_i^\omega - \phi^\omega|)$ respectively. Gossip among peers increases actual social disapproval for social norm violation and non-conformity proportionally. Hence, we write $v(\phi^\omega) = \tilde{v}(\phi^\omega)(1+\delta)$ and $k(|n_i^\omega - \phi^\omega|) = \tilde{k}(|n_i^\omega - \phi^\omega|)(1+\delta)$. Throughout, we assume that $\tilde{v}(\phi^\omega)$ and $\tilde{k}(|n_i^\omega - \phi^\omega|)$ are continuous, differentiable and invertible.

The expression of social disapproval for hypocrisy requires observers of i's action are aware of her respective personal norm and vice versa. Either the observers have previously observed it, or others in society shared the information with them. Therefore, expressed social disapproval for hypocrisy $\tilde{h}$ is linked to the level of gossip δ. Moreover, some actual social disapproval for hypocrisy only arises from pooling information through gossip and

[9]Note that feelings of social disapproval can also be triggered internally. Individuals know the distribution of norms in society and form expectations about the personal norms of their peers. An individual that believes her peers disapprove of her experiences negative feelings. Changing the setup accordingly alters the underlying story but not the formal analysis.

drawing conclusions therefrom.[10] Thus, social disapproval for hypocrisy is disproportionally greater than the expressed one such that $h > \tilde{h}(1+\delta)$.

Definition 2.4 (Expressed Social Approval)**.** $\tilde{s}_i^\omega(a_i^\omega, \phi^\omega, n_i^\omega) = (a_i^\omega - 1)[\tilde{v}(\phi^\omega) + n_i^\omega \tilde{h}] + \tilde{k}(|n_i^\omega - \phi^\omega|)$.

3 Behavior

This section discusses the evolution of behavior. Behavior evolves at the individual level and does so significantly faster than norms and preferences.

3.1 Evolutionary Framework

The evolution of behavior is driven by utility. An individual i's utility for situation ω depends on her material payoff m_i^ω, self-approval p_i^ω, and the social approval expressed towards her $\tilde{s}_i^\omega$. The degree to which these components determine utility depends on her preference type $\theta_i = (\theta_{s_i}, \theta_{p_i}) \in \Theta$, where θ_{s_i} and θ_{p_i} capture the impact of social and self-approval respectively.

Definition 3.1 (Utility)**.** $u_i^\omega(a_i^\omega, n_i^\omega, \psi^\omega, \phi^\omega, \theta_i) = m_i^\omega(a_i^\omega, \psi^\omega) + \theta_{s_i}\tilde{s}_i^\omega(a_i^\omega, n_i^\omega, \phi^\omega) + \theta_{p_i} p(a_i^\omega, n_i^\omega)$.

Let $\sigma_{n,\theta}^\omega$ be the share of individuals with personal norm $n \in \{0,1\}$ and preference type θ that contribute to the public good. Moreover, $\sigma_n^\omega = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \sigma_{n,\theta}^\omega$ indicates the share of cooperators among individuals with personal norm $n_i^\omega = n$. Hence, we can write the cooperation level as $\psi^\omega = \phi^\omega \sigma_1^\omega + (1-\phi^\omega)\sigma_0^\omega$. The vector σ^ω consists of all vectors $(\sigma_{1,\theta}^\omega, \sigma_{0,\theta}^\omega)$ and thus describes the complete distribution of behavior in ω. Lastly, the vector σ consists of behavioral distributions σ^ω for each situation $\omega \in \Omega$.

For each situation ω, an individual i has a behavioral routine which specifies her action a_i^ω. Once in a while, an individual questions her current behavioral routine and revises it. The individual switches routines with positive probability if this yields utility gains at the

[10] For example, consider two observers of i in two separate instances so that one only observed a_i^ω and the other only n_i^ω. Gossip between these two individuals potentially leads to social disapproval of i if it reveals that i behaves inconsistently with her own personal norms. However, none of the two observers previously expressed disapproval for hypocrisy towards i.

current environment. Formally, we can describe this evolutionary process using pairwise comparison dynamics (Sandholm 2010).

Definition 3.2 (Behavioral Dynamics)**.**

$$\begin{aligned}\dot{\sigma}^{\omega}_{n,\theta} &= f(u^{\omega}(1,n,\psi^{\omega},\phi^{\omega},\theta) - u^{\omega}(0,n,\psi^{\omega},\phi^{\omega},\theta), \sigma^{\omega}_{n,\theta}) \\ &= f(\Delta m^{\omega}(\psi^{\omega}) + \theta_s \tilde{v}(\phi^{\omega}) + n(\theta_s \tilde{v} + \theta_p), \sigma^{\omega}_{n,\theta}) \, \forall n \in \{0,1\} \text{ and } \theta \in \text{supp}(\lambda),\end{aligned}$$

where f satisfies:

1. f is Lipschitz continuous,
2. $f(\Delta u^{\omega}_{n,\theta}, \sigma^{\omega}_{n,\theta}) > 0 \Leftrightarrow (\Delta u^{\omega}_{n,\theta} > 0 \wedge \sigma^{\omega}_{n,\theta} \neq 1)$, and
3. $f(\Delta u^{\omega}_{n,\theta}, \sigma^{\omega}_{n,\theta}) < 0 \Leftrightarrow (\Delta u^{\omega}_{n,\theta} < 0 \wedge \sigma^{\omega}_{n,\theta} \neq 0)$.

As utility in ω only depends on the action that is executed in that particular situation, we can investigate the evolution of behavior for each situation in isolation. Throughout, we employ the following equilibrium notion.

Definition 3.3 (Behavioral Equilibrium)**.** A connected, non-empty and closed set Σ^{ω} of behavioral distributions σ^{ω} is a *behavioral equilibrium* if it is asymptotically stable in 3.2 and there is no $\tilde{\Sigma}^{\omega} \subset \Sigma^{\omega}$ such that $\tilde{\Sigma}^{\omega}$ is asymptotically stable in dynamics 3.2. We indicate a behavioral equilibrium by $\Sigma^{\omega *}$ or, if it is a singleton, $\sigma^{\omega *}$.

3.2 Equilibrium Analysis

The following proposition summarizes the main result of this section.

Proposition 3.1. *For each $\omega \in \Omega$, $\lambda \in [0,1]^{|\Theta|}$, and $\phi^{\omega} \in [0,1]$, there is a unique behavioral equilibrium $\Sigma^{\omega *}$ s.t.*

1. *$\sigma^{\omega} \in \Sigma^{\omega *} \Leftrightarrow \sigma^{\omega}$ is a Nash equilibrium and*
2. *for all $\hat{\sigma}^{\omega}, \check{\sigma}^{\omega} \in \Sigma^{\omega *}$, $\phi^{\omega}\hat{\sigma}^{\omega}_1 + (1-\phi^{\omega})\hat{\sigma}^{\omega}_0 = \phi^{\omega}\check{\sigma}^{\omega}_1 + (1-\phi^{\omega})\check{\sigma}^{\omega}_0$.*

Proof: proposition 3.1 is a corralory of lemmas B.1 and B.3 in appendix B.1.

The above states that there always exists a unique behavioral equilibrium $\Sigma^{\omega*}$ that society unambiguously coordinates into. This behavioral equilibrium coincides with the set of all Nash equilibria. Moreover, all behavioral distributions $\sigma^{\omega} \in \Sigma^{\omega*}$ yield the same cooperation share $\psi^{\omega*}$. At times, we may write the behavioral equilibrium and the equilibrium cooperation level explicitly as functions of the social norm ϕ^{ω} and preference distribution λ: $\Sigma^{\omega*}(\phi^{\omega}, \lambda)$ and $\psi^{\omega*}(\phi^{\omega}, \lambda)$.

To illustrate the intuition behind the above result, we start by discussing a society that is homogeneous regarding approval preferences. In this case, the behavioral equilibrium is a singleton (see lemma B.8 in appendix B.1). Consider the graphical illustration in figure 1. The costs of contribution at any cooperation level ψ^{ω} are $-\Delta m^{\omega}(\psi^{\omega})$. By definition 3.2, $-\Delta m^{\omega}(\psi^{\omega})$ is strictly increasing. Next, consider the function $NU^{\omega}(\psi^{\omega}, \phi^{\omega})$. It sorts all individuals' social and self-approval gains from cooperation in descending order. The first ϕ^{ω} individuals hold the cooperation norm and thus avoid social disapproval from social norm violation $\theta_s \tilde{v}(\phi^{\omega})$, social disapproval from hypocrisy $\theta_s \tilde{h}$, and self-disapproval θ_p when cooperating. The remaining $1 - \phi^{\omega}$ individuals do not hold the cooperation norm and thus only avoid social disapproval from social norm violation $\theta_s \tilde{v}(\phi^{\omega})$. $NU^{\omega}(\psi^{\omega}, \phi^{\omega})$ is a decreasing function by construction.

If at some cooperation level ψ^{ω}, $NU^{\omega}(\psi^{\omega}, \phi^{\omega})$ lies above $-\Delta m^{\omega}(\psi^{\omega})$, then some individual who is currently not cooperating prefers to do so. By changing her behavioral routine accordingly, she would obtain greater norm-based utility benefits than material costs. Similarly, if $NU^{\omega}(\psi^{\omega}, \phi^{\omega})$ lies below $-\Delta m^{\omega}(\psi^{\omega})$, then some individual who currently cooperates prefers to defect. The behavioral equilibrium is given by the unique intersection of both curves or, if no intersection exists, at one of the extremes. Subplot 1 (a) shows an equilibrium in which all individuals who hold the cooperation norm strictly prefer to cooperate, and all individuals who do not hold the cooperation norm strictly prefer not to cooperate. Figure 1 (b) provides an example in which some individuals who do not hold the cooperation norm are sufficiently motivated to cooperate. The material costs of contribution equal social disapproval from social norm violation in a manner that all norm non-holders are indifferent.

Next, we generalize the above discussion to the case of heterogeneous approval preferences. As with homogeneous preferences, we can sort all individuals according to their utility

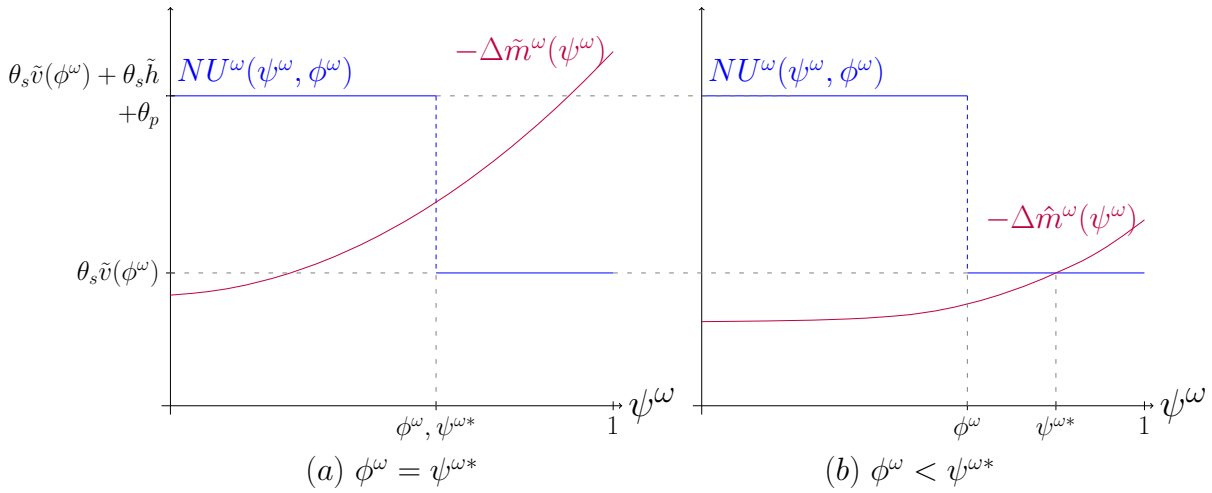


Figure 1: behavioral equilibrium.

gains from cooperation, which yields a decreasing function. The unique equilibrium share of cooperators $\psi^{\omega*}$ is at this function's intersection with the material costs of contribution $-\Delta m^\omega(\psi^\omega)$. However, the behavioral equilibrium is no longer necessarily a unique stable point, but possibly a connected, non-empty, and stable set of rest points $\Sigma^{\omega*}(\phi^\omega, \lambda)$. This may occur if at least two sub-groups of individuals $\hat{\mathcal{I}} = \{i \in \mathcal{I} : n_i = \hat{n} \wedge \theta_i = \hat{\theta}\}$ and $\check{\mathcal{I}} := \{i \in \mathcal{I} : n_i = \check{n} \wedge \theta_i = \check{\theta}\}$ are indifferent at the equilibrium costs of contribution, $\hat{\theta}_s\tilde{v}(\phi^\omega) + \hat{n}\hat{\theta}_s\tilde{h} + \hat{n}\hat{\theta}_p = -\Delta m^\omega(\psi^{\omega*}) = \check{\theta}_s\tilde{v}(\phi^\omega) + \check{n}\check{\theta}_s\tilde{h} + \check{n}\check{\theta}_p$. If so, there may exist infinitely many Nash equilibria σ^ω that all exhibit varying sub-group cooperation levels $\sigma^\omega_{\hat{\theta},\hat{n}}$ and $\sigma^\omega_{\check{\theta},\check{n}}$, but the same cooperation level $\psi^{\omega*}$.

We can utilize the results of this section to discuss how changes in different variables affect equilibrium behavior. For illustration purposes, the following discussion focuses on a society that is homogeneous regarding approval preferences. Therefore, consider figure 1. An increase in the social norm ϕ^ω implies that more individuals hold the cooperation norm. Therefore, social disapproval for social norm violation also increases. Graphically, both horizontal segments of $NU(\psi^\omega, \phi^\omega)$ shift upwards, and their vertical connection moves to the right. Consequently, the equilibrium level of cooperation must rise. Moreover, the equilibrium level of cooperation also increases in preferences for self-approval θ_p, preferences for social approval θ_s, expressed social disapproval from social norm violation $\tilde{v}$, and expressed social disapproval from hypocrisy $\tilde{h}$. These insights easily carry over to the general case of

heterogeneous approval preferences.

4 Norms

Next, we turn to norm evolution. Norms evolve through cultural transmission and do so significantly slower than behavior. Therefore, we assume that society always reaches a behavioral equilibrium before further changes in the norm distribution occur. Moreover, we assume that an individual's personal norm for one situation does not affect the internalization process of norms in other situations. Consequently, we can investigate norm evolution independently in each situation.

4.1 Evolutionary Framework

Generally, the personal norms of culturally successful individuals spread in society. We assume that cultural success depends on material factors (e.g., income, occupational prestige) and social ones (e.g., social reputation, respect). Thus, material payoff and social approval co-determine the cultural fitness that drives norm evolution.

Definition 4.1 (Cultural Fitness)**.** $c_i^\omega(a_i^\omega, n_i^\omega, \psi^\omega, \phi^\omega) = m_i^\omega(a_i^\omega, \psi^\omega) + \gamma s_i^\omega(a_i^\omega, n_i^\omega, \phi^\omega)$, where $0 < \gamma$ is the weight of social approval on cultural fitness.

Following the existing literature, we assume cultural transmission of norms mainly occurs through horizontal (peer interactions) and oblique transmission (socialization institutions). First, individuals are more likely to copy the cultural traits of culturally successful peers (Joseph Henrich and Gil-White 2001). Second, access to specific social networks as well as financial means favors the chances of acquiring privileged cultural positions (e.g., teachers, politicians), which, in turn, increases the impact on the opinion formation process of others (Bowles and Gintis 1998). Access to certain social networks is often denied if an individual is subject to social disapproval (e.g., Cinyabuguma et al. 2005; Traxler and Spichtig 2011). As norms evolve based on learning through socialization (rather than self-improvement), we assume that norm internalization occurs independently of an individual's approval preferences. Hence, norms are independently distributed across all sub-populations of preference

types.[11] Formally, we can best describe norm evolution using imitative dynamics (see Sandholm 2010). Therefore, we employ the well-studied replicator dynamics.[12]

Definition 4.2 (Norm Dynamics)**.**

$$\begin{aligned}\dot{\phi}^{\omega} &= \phi^{\omega}(1-\phi^{\omega})(C_1^{\omega}(\sigma^{\omega},\phi^{\omega}) - C_0^{\omega}(\sigma^{\omega},\phi^{\omega})) \\ &= \phi^{\omega}(1-\phi^{\omega})((\sigma_1^{\omega}-\sigma_0^{\omega})(\gamma v(\phi^{\omega}) + \Delta m^{\omega}(\psi^{\omega})) - \gamma(1-\sigma_1^{\omega})h + \gamma\Delta k(\phi^{\omega})),\end{aligned}$$

where $\Delta k(\phi^{\omega}) = k(|1-\phi^{\omega}|) - k(|0-\phi^{\omega}|)$ and $C_n^{\omega}(\sigma^{\omega},\phi^{\omega}) = \sigma_n^{\omega}c^{\omega}(1,n,\psi^{\omega},\phi^{\omega}) + (1-\sigma_n^{\omega})c^{\omega}(0,n,\psi^{\omega},\phi^{\omega})$ is the average cultural fitness of all individuals with personal norm $n \in \{0,1\}$ from situation ω.

Similar to behavior, we are interested in minimal asymptotically stable sets of rest points. Therefore, we employ the following equilibrium notion.

Definition 4.3 (Cultural Equilibrium)**.** A connected, non-empty and closed set Φ^{ω} of social norms ϕ^{ω} is a *cultural equilibrium* if it is asymptotically stable in 4.2 and there is no $\tilde{\Phi}^{\omega} \subset \Phi^{\omega}$ such that $\tilde{\Phi}^{\omega}$ is asymptotically stable in dynamics 4.2. We indicate a cultural equilibrium by $\Phi^{\omega *}$ or, if it is a singleton, $\phi^{\omega *}$.

4.2 Equilibrium Analysis

This section discusses the results on norm evolution that are most relevant for our further analysis.[13] First, we present cultural equilibria that exist if approval preferences are homogeneous. Thereafter, we generalize the results to the heterogeneous preference case and present some additional insights.

[11] A consequence thereof is that any equilibrium cooperation share of norm non-holders $\sigma_0^{\omega *}$ never exceeds that of norm holders $\sigma_1^{\omega *}$ (see lemma B.6 in appendix B.1).

[12] Following Sandholm (2010), we can easily show that our results also hold for a variety of other populations dynamics.

[13] Appendix A.2 discusses additional cultural equilibria that may exist under homogeneous approval preferences.

4.2.1 Homogeneous Approval Preferences

This section presents asymptotically stable points of dynamics 4.2 that exist under homogeneous approval preferences. Therefore, we assume throughout this section that each individual is of preference type θ.

Proposition 4.1 (Cultural Equilibrium of no Social Norm under Homogeneous Preferences)**.** *Suppose society is homogeneous regarding approval preferences. For any situation $\omega \in \Omega$, the social norm $\phi^\omega = 0$ is always a cultural equilibrium.*

Proof: The proposition is a special case of proposition 4.4 and thus needs no separate proof. We present it here mainly for expositional reasons.

The first cultural equilibrium corresponds to the absent social norm $\phi^{\omega*} = 0$, where no individual holds the cooperation norm. All individuals have neither personal nor social incentives to cooperate and, thus, defect, $\psi^{\omega*}(0, \lambda) = 0$. Such a cultural equilibrium always exists. Starting from $\phi^{\omega*} = 0$, assume that a small group of norm holders appears. As the new social norm ϕ^ω remains very close to zero, the norm holders are subject to high social disapproval from non-conformity. Moreover, the internalization of the cooperation norm may induce them to either change their behavioral routine to cooperation or keep defecting. In the former case, the norm holders incur material costs but avoid social disapproval from social norm violation. However, the avoided social disapproval barely impacts differences in cultural fitness as the social norm ϕ^ω is close to zero. In the latter case, all individuals obtain the same material payoff and social disapproval from social norm violation. However, the norm holders behave hypocritically, negatively impacting their cultural fitness. Consequently, the norm holders obtain lower cultural fitness on average in both cases, inducing a return to $\phi^{\omega*} = 0$.

Proposition 4.2 (Cultural Equilibrium of a Perfect Social Norm under Homogeneous Preferences)**.** *Suppose the distribution of preferences λ is such that society is homogeneous with preference type θ. For any situation $\omega \in \Omega$, the social norm $\phi^\omega = 1$ is a cultural equilibrium if $\psi^{\omega*}(1, \lambda)(\gamma v(1) + \Delta m^\omega(\psi^{\omega*}(1, \lambda))) - \gamma(1 - \psi^{\omega*}(1, \lambda))h + \gamma\Delta k(1) > 0$.*

Proof: The proposition is a special case of proposition 4.5 and thus needs no separate proof. We present it here mainly for expositional reasons.

Proposition 4.2 presents a cultural equilibrium where all individuals hold the cooperation norm so that the social norm is perfect, $\phi^{\omega*} = 1$. Either all individuals cooperate, $\psi^{\omega*}(1, \lambda) = 1$, or some do not, $\psi^{\omega*}(1, \lambda) < 1$. First, we investigate the case where all individuals cooperate. Consider a cultural mutation that leads to some individuals abandoning the cooperation norm. The new social norm ϕ^{ω} remains close to one. If the contribution costs are small, individuals who abandon the cooperation norm do not change their behavior. All individuals cooperate and incur the same social disapproval from norm violation $v(\phi^{\omega})$ and costs of contribution $-\Delta m^{\omega}(\psi^{\omega}(1, \lambda))$. Hypocrisy does not occur. As the social norm after mutation is close to one, the norm non-holders experience greater social disapproval from non-conformity, $\gamma\Delta k(\phi^{\omega}) > 0$. Hence, they obtain lower cultural fitness on average. Alternatively, individuals who abandon the cooperation norm are no longer sufficiently motivated to cooperate. The decrease in the social norm also changes behavioral incentives for individuals who still hold the cooperation norm due to a change in social disapproval from social norm violation. However, since the change in the social norm is small, cooperation incentives only change slightly, and the respective share of cooperators $\sigma_1^{\omega*}$ remains very close to one. Consequently, hypocrisy is negligible. If social disapproval from social norm violation and non-conformity on cultural fitness is greater than the costs of contribution at the perfect social norm, $\gamma v(1) + \gamma\Delta k(1) > -\Delta m^{\omega}(1)$, then this also holds for marginally smaller social norms ϕ^{ω}, $\gamma v(\phi^{\omega}) + \gamma\Delta k(\phi^{\omega}) > -\Delta m^{\omega}(\psi^{\omega*}(\phi^{\omega}, \lambda))$. Hence, the defecting norm non-holders obtain lower cultural fitness on average.

Next, we investigate the case of partial cooperation $\psi^{\omega*}(1, \lambda) < 1$ at the perfect social norm. By the same reasoning as above, the post-mutation share of cooperators among norm holders $\sigma_1^{\omega*}$ remains close to the pre-mutation level of cooperation $\psi^{\omega*}(1, \lambda)$. The individuals who abandon the cooperation norm must strictly prefer not to cooperate, $\sigma_0^{\omega*} = 0$, since even some individuals who hold the cooperation norm prefer not to. As the post-mutation social norm ϕ is close to one, social disapproval form social norm violation $v(\phi^{\omega})$ and non-conformity $\Delta k(\phi^{\omega})$ remain close to $\gamma v(1)$ and $\gamma\Delta k(1)$ respectively. The condition stated in proposition 4.2, $\psi^{\omega*}(1, \lambda)(\gamma v(1) + \Delta m^{\omega}(\psi^{\omega*}(1, \lambda))) - \gamma(1 - \psi^{\omega*}(1, \lambda))h + \gamma\Delta k(1) > 0$, then

implies that at the new social norm ϕ^{ω}, average differences in social disapproval from non-conformity and social norm violation outweigh material payoff differences and average social disapproval from hypocrisy, $\sigma_1^{\omega*}(\gamma v(\phi^{\omega}) + \Delta m^{\omega}(\psi^{\omega*}(\phi^{\omega}, \lambda)) - \gamma(1 - \sigma_1^{\omega*})h + \gamma \Delta k(\phi^{\omega}) > 0$. The norm holders obtain greater cultural fitness than the norm non-holders.

Thus, the stated condition ensures that, on average, the norm holders have greater cultural fitness if some small group of norm non-holders arises. Therefore, society returns to the perfect norm after some cultural mutation, implying that it is a cultural equilibrium.

Proposition 4.3 (Cultural Equilibrium of an Imperfect Social Norm under Homogeneous Preferences)**.** *Suppose society is homogeneous regarding approval preference with preference type θ. For any situation $\omega \in \Omega$, the social norm $\phi^{\omega} \in (0, 1)$ is a cultural equilibrium if*

1. $\theta_s \tilde{v}(\phi^{\omega}) < -\Delta m^{\omega}(\phi^{\omega}) < \theta_s \tilde{v}(\phi^{\omega}) + \theta_s \tilde{h} + \theta_p$,
2. $\gamma(v(\phi^{\omega}) + \Delta k(\phi^{\omega})) = -\Delta m^{\omega}(\phi^{\omega})$, *and*
3. $\gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi^{\omega}} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi^{\omega}}) < -\frac{\mathrm{d}\Delta m^{\omega}(x)}{\mathrm{d}x}|_{x=\phi^{\omega}}$.

Proof: The proposition is a special case of proposition 4.7 and thus needs no separate proof. We present it here mainly for expositional reasons.

Proposition 4.3 describes an equilibrium with an imperfect social norm $\phi^{\omega*} \in (0, 1)$. Figure 2 presents a graphical illustration thereof. At the respective imperfect social norm, individuals cooperate if and only if they holds the cooperation norm. For norm evolution to be at rest, the average cultural fitness of the norm holders must equal that of the norm non-holders. This is satisfied if the costs of contribution $-\Delta m^{\omega}(\phi^{\omega*})$ equal the differences in social disapproval from non-conformity and social norm violation on cultural fitness $\gamma(v(\phi^{\omega*}) + \Delta k(\phi^{\omega*}))$. Lastly, suppose some individuals randomly internalize (abandon) the cooperation norm. In that case, the average cultural fitness of the norm holders must fall below (above) that of the norm non-holders for society to return to the initial state. This holds if an increase (decrease) in $\phi^{\omega*}$ leads to a more drastic increase (decrease) in the material costs of contribution than the difference in social approval. Graphically, $-\Delta m^{\omega}(\phi^{\omega})$ intersects $\gamma(v(\phi^{\omega}) + \Delta k(\phi^{\omega}))$ from below at $\phi^{\omega*}$.

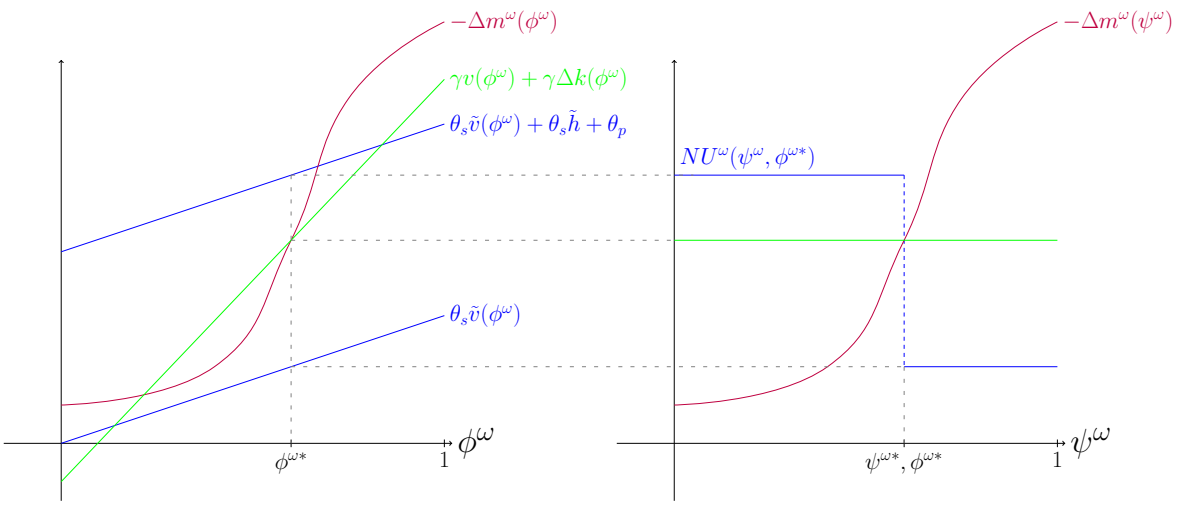


Figure 2: cultural equilibrium of an imperfect social norm.

4.2.2 Heterogeneous Approval Preferences

We continue by generalizing the results on norm evolution to the case of heterogeneous approval preferences. Moreover, we introduce additional results, which will prove helpful in later stages of the analysis.

Proposition 4.4 (Cultural Equilibrium of no Social Norm)**.** *For all* $\omega \in \Omega$ *and* $\lambda \in [0,1]^{|\Theta|}$, $\phi^{\omega} = 0$ *is a cultural equilibrium.*

Proof: See appendix B.2.

Proposition 4.4 states that under heterogeneous preferences, the cultural equilibrium of no social norm always exists. The underlying reasoning coincides with that for homogeneous approval preferences.

Proposition 4.5 (Cultural Equilibrium of a Perfect Social Norm)**.** *For all* $\lambda \in [0,1]^{|\Theta|}$ *and* $\omega \in \Omega$, $\phi^{\omega} = 1$ *is a cultural equilibrium if*

1. $\psi^{\omega *}(1,\lambda)(\gamma v(1) + \Delta m^{\omega}(\psi^{\omega *}(1,\lambda))) - \gamma(1 - \psi^{\omega *}(1,\lambda))h + \gamma \Delta k(1) > 0$ *and*
2. *(a)* $\theta_s \tilde{v}(1) < -\Delta m^{\omega}(\psi^{\omega *}(1,\lambda))$ *for all* $\theta \in \text{supp}(\lambda)$ *or*

   (b) $\Delta k(1) > (1 - \psi^{\omega *}(1,\lambda))h$.

Proof: See appendix B.2.

Proposition 4.5 states that under heterogeneous preferences, the perfect social norm is a cultural equilibrium if, additionally to condition 1 which is the same as under homogeneous approval preferences (see proposition 4.2), either (a) an individual would prefer to defect if she was holding the cooperation norm or (b) social disapproval for non-conformity outweighs average social disapproval for hypocrisy. Note that condition 2 is always satisfied if cooperation is full, $\psi^{\omega*}(1,\lambda) = 1$. Hence, condition 1 ensures stability of a perfect social norm that induces full cooperation. The underlying reasoning closely follows the corresponding homogeneous preference case.

If cooperation at $\phi^{\omega} = 1$ is partial, $\psi^{\omega*}(1,\lambda) < 1$, condition 1 alone is insufficient to ensure stability. Recall that the stability argument under homogeneous preferences builds on the fact that after some small cultural mutation to $\phi^{\omega} < 1$, all individuals who do not hold the cooperation norm defect, $\sigma_0^{\omega*} = 0$. Under heterogeneous preferences, this is not necessarily the case. Condition 2a ensures that it is. If it holds, the stability argument of the perfect social norm again closely follows that for homogeneous approval preferences.

Alternatively, condition 2b implies that after some cultural mutation to $\phi^{\omega} < 1$, social disapproval for non-conformity $\Delta k(\phi^{\omega})$ outweighs social disapproval for hypocrisy $(1-\sigma_1^{\omega*})h$. If some norm non-holders cooperate at ϕ^{ω}, then the difference in norm population behavior decreases. Differences in social disapproval from norm social violation and material costs become less pronounced, $(\sigma_1^{\omega*} - \sigma_0^{\omega*})(\gamma v(\phi^{\omega}) + \Delta m^{\omega}(\phi^{\omega})) < \sigma_1^{\omega*}(\gamma v(\phi^{\omega}) + \Delta m^{\omega}(\phi^{\omega}))$. The relative impact of social disapproval from hypocrisy and non-conformity on differences in cultural fitness rises. This ensures that the norm holder obtain greater cultural fitness than the norm non-holders on average, $(\sigma_1^{\omega*}-\sigma_0^{\omega*})(\gamma v(\phi^{\omega})+\Delta m^{\omega}(\phi^{\omega}))+\gamma(1-\sigma_1^{\omega*})h+\Delta k(\phi^{\omega}) > 0$. Consequently, $\phi^{\omega*} = 1$ is a cultural equilibrium.

Proposition 4.6 (Robustness of a Cultural Equilibrium of a Perfect Social Norm)**.** *Consider any $\omega \in \Omega$ and $\lambda \in [0,1]^{|\Theta|}$ for which a cultural equilibirum of a perfect social norm $\phi^{\omega*} = 1$ of proposition 4.5 exists. There is U of λ s.t. $\phi^{\omega} = 1$ is a cultural equilibrium at any $\hat{\lambda} \in U$.*

Proof: See appendix B.2.

Proposition 4.6 establishes that a cultural equilibrium $\phi^{\omega*} = 1$ of proposition 4.5 that exists at preference distribution λ is also a cultural equilibrium at any other preference

distribution $\hat{\lambda}$, if $\hat{\lambda}$ is sufficiently close to λ. The intuition behind this result is as follows. At preference distribution λ and any social norm ϕ^ω in the neighborhood of $\phi^{\omega*} = 1$, the norm holders obtain strictly greater cultural fitness than the norm non-holders, $(\sigma_1^{\omega*} - \sigma_0^{\omega*})(\gamma v(\phi^\omega) + \Delta m^\omega(\phi^\omega)) + \gamma(1 - \sigma_1^{\omega*})h + \Delta k(\phi^\omega) > 0$. If the preference distribution $\hat{\lambda}$ is close to λ, then differences in norm population behavior $(\sigma_1^{\omega*}, \sigma_0^{\omega*})$ at any social norm ϕ^ω at both preference distributions are small. Thus, the norm holders still obtain greater cultural fitness than the norm non-holders at $\hat{\lambda}$ in the neighborhood of $\phi^{\omega*} = 1$. It follows that the perfect social norm is also a cultural equilibrium at $\hat{\lambda}$.

Proposition 4.7 (Cultural Equilibrium of an Imperfect Social Norm)**.** *For all $\omega \in \Omega$, $\lambda \in [0,1]^{|\Theta|}$, and $\phi^\omega \in (0,1)$, ϕ^ω is a cultural equilibrium if*

1. $\theta_s \tilde{v}(\phi^\omega) < -\Delta m^\omega(\phi^\omega) < \theta_s \tilde{v}(\phi^\omega) + \theta_s \tilde{h} + \theta_p \ \forall \theta \in \text{supp}(\lambda)$,
2. $\gamma(v(\phi^\omega) + \Delta k(\phi^\omega)) = -\Delta m^\omega(\phi^\omega)$, *and*
3. $\gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi^\omega} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi^\omega}) < -\frac{\mathrm{d}\Delta m^\omega(x)}{\mathrm{d}x}|_{x=\phi^\omega}$.

Proof: See appendix B.2.

Next, we investigate a cultural equilibrium of an imperfect social norm $\phi^{\omega*}$. If at the social norm $\phi^{\omega*}$ all norm non-holders strictly prefer to defect and norm holders strictly prefer to cooperate, $\theta_s \tilde{v}(\phi^{\omega*}) < -\Delta m^\omega(\phi^{\omega*}) < \theta_s \tilde{v}(\phi^{\omega*}) + \theta_s \tilde{h} + \theta_p \ \forall \theta \in \text{supp}(\lambda)$, then norm population behavior is $(\sigma_1^{\omega*}, \sigma_0^{\omega*}) = (1, 0)$ in some neighborhood of $\phi^{\omega*}$. By the same reasoning as under homogeneous preferences, the remaining two conditions of proposition 4.7 imply that $\phi^{\omega*}$ is a cultural equilibrium. Throughout the following, we consider some situation $\omega \in \Omega$ and preference distribution $\lambda \in [0,1]^{|\Theta|}$ for which a cultural equilibrium of an imperfect social norm $\phi^{\omega*} \in (0,1)$ exists.

Proposition 4.8 (Robustness of a Cultural Equilibrium of an Imperfect Social Norm)**.** *Consider any $\omega \in \Omega$ and $\lambda \in [0,1]^{|\Theta|}$ for which a cultural equilibrium of an imperfect social norm $\phi^{\omega*} \in (0,1)$ of proposition 4.7 exists. For all $\epsilon > 0$, there is U of λ s.t. at any $\hat{\lambda} \in U$, there exists a cultural equilibrium $\hat{\Phi}^{\omega*} \subset (\phi^{\omega*} - \epsilon, \phi^{\omega*} + \epsilon)$.*

Proof: Follows from proposition B.1 in appendix B.2.

Proposition 4.8 states that if there exists a cultural equilibrium of an imperfect social norm $\phi^{\omega*}$ at preference distribution λ, then for all preference distributions $\hat{\lambda}$ close to λ there exists a cultural equilibrium $\hat{\Phi}^{\omega*}$ close to $\phi^{\omega*}$. Figure 3 illustrates the underlying intuition of this result graphically. At preference distribution λ, the right intersection of the two solid lines constitutes the cultural equilibrium $\phi^{\omega*}$. Consider some preference distribution $\hat{\lambda}$ that differs only slightly from λ. Let ϵ_1 be the share of individuals that prefer to defect at $\phi^{\omega*}$ and some neighborhood if they hold the cooperation norm. Analogously, let ϵ_0 be the share of individuals that prefer to cooperate if they do not hold the cooperation norm. If $\hat{\lambda}$ is close to λ, ϵ_0 and ϵ_1 are close to zero, and equilibrium behavior in some neighborhood of $\phi^{\omega*}$ is $\sigma_1^{\omega*} = 1 - \epsilon_1$, $\sigma_0^{\omega*} = \epsilon_0$, and $\psi^{\omega*} = \phi^{\omega*}(1 - \epsilon_1) + (1\phi^{\omega*})\epsilon_0$. Substituting this into dynamics 4.2 yields that norm evolution is at rest if $(1 - \epsilon_1 - \epsilon_0)(\gamma v(\phi^{\omega}) + \Delta m^{\omega}(\phi^{\omega}(1 - \epsilon_1) + (1 - \phi^{\omega})\epsilon_0)) - \gamma\epsilon_1 h + \gamma\Delta k(\phi^{\omega}) = 0$. The intersection of the two dotted lines at $\hat{\phi}^{\omega*}$ in figure 3 represents such a rest point. It is asymptotically stable since the cost curve (red) intersects the social approval curve (green) from below. The closer $\hat{\lambda}$ to λ, the closer the dotted to the solid lines, and, consequently, $\hat{\phi}^{\omega*}$ to $\phi^{\omega*}$.

Proposition 4.9 (Instability of an Imperfect Social Norm)**.** *Consider any $\omega \in \Omega$ and $\lambda \in [0,1]^{|\Theta|}$ for which a cultural equilibrium of an imperfect social norm $\phi^{\omega*} \in (0,1)$ of proposition 4.7 exists. Let $\hat{\lambda} \in [0,1]^{|\Theta|}$ satisfy*

- $\theta_s\tilde{v}(\phi^{\omega*}) \leq -\Delta m^{\omega}(\phi^{\omega*}) < \theta_s\tilde{v}(\phi^{\omega*}) + \theta_s\tilde{h} + \theta_p$ *for all* $\theta \in \text{supp}(\hat{\lambda})$ *and*
- $\theta_s\tilde{v}(\phi^{\omega*}) = -\Delta m^{\omega}(\phi^{\omega*})$ *for some* $\theta \in \text{supp}(\hat{\lambda})$.

There is no cultural equilibrium $\hat{\Phi}^{\omega*}$ at $\hat{\lambda}$ s.t. $\phi^{\omega*} \in \hat{\Phi}^{\omega*}$ if

1. $\phi^{\omega*} < \frac{1}{2}$ *and*
2. $\gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}}) > \frac{-\Delta m^{\omega}(\phi^{\omega*})}{\tilde{v}(\phi^{\omega*})} \times \frac{\mathrm{d}\tilde{v}(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}}$.

Proof: See appendix B.2.

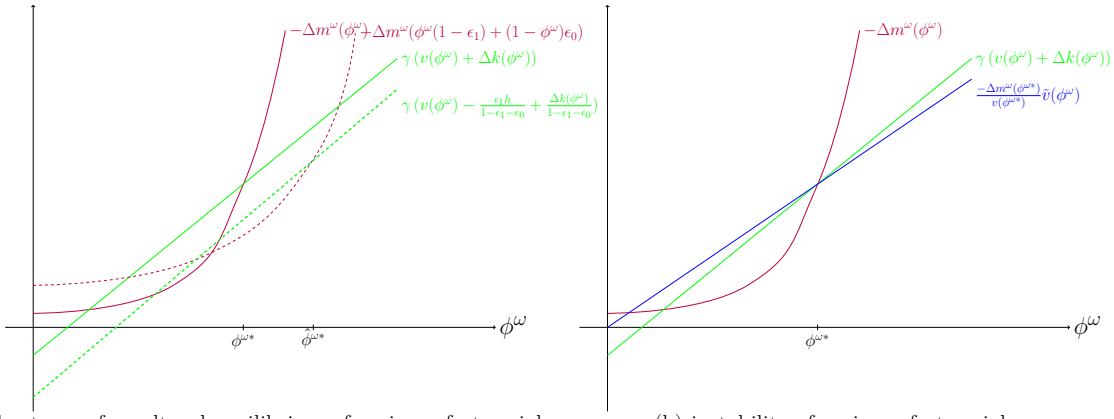
Finally, proposition 4.9 states that a cultural equilibrium of an imperfect social norm $\phi^{\omega*}$ at preference distribution λ may not be (part of) a cultural equilibrium at another preference

distribution $\hat{\lambda}$, even if $\hat{\lambda}$ differs from λ only in that some norm non-holders are indifferent between both behavioral routines at $\phi^{\omega*}$. Hence, although preference distributions λ and $\hat{\lambda}$ both induce equilibrium behavior $(\sigma_1^{\omega*}, \sigma_0^{\omega*}) = (1, 0)$ at $\phi^{\omega*}$, the respective social norm is a cultural equilibrium at preference distribution λ but not at preference distribution $\hat{\lambda}$. The underlying reason is that norm population behavior in the neighborhood of the social norm $\phi^{\omega*}$ may differ for both preference distributions. Below, we discuss why the proposition holds in more detail.

Consider the preference type $\bar{\theta} \in \text{supp}(\hat{\lambda})$ that induces an individual to be indifferent between both behavioral routines when not holding the cooperation norm, $\bar{\theta}_s \tilde{v}(\phi^{\omega*}) = -\Delta m^{\omega}(\phi^{\omega*}) \Rightarrow \bar{\theta}_s = \frac{-\Delta m^{\omega}(\phi^{\omega*})}{\tilde{v}(\phi^{\omega*})}$. Both preference distributions λ and $\hat{\lambda}$ yield norm population behavior $(\sigma_1^{\omega*}, \sigma_0^{\omega*}) = (1, 0)$ at social norm $\phi^{\omega*}$. Since $\phi^{\omega*}$ is a rest point of norm dynamics 4.2 at λ, it must thus also be a rest point at $\hat{\lambda}$. Consider some cultural mutation to a weaker social norm $\phi^{\omega} < \phi^{\omega*}$.

At preference distribution λ, all individuals cooperate at the new social norm ϕ^{ω} if and only if they hold the cooperation norm, $(\sigma_1^{\omega*}, \sigma_0^{\omega*}) = (1, 0)$. Moreover, from condition 3 of proposition 4.7 we know that at the new social norm ϕ^{ω}, social disapproval for social norm violation and non-conformity on cultural fitness exceed equilibrium cooperation costs, $\gamma\left(v(\phi^{\omega}) + \Delta k(\phi^{\omega})\right) > -\Delta m^{\omega}(\phi^{\omega*})$. As discussed before, the social norm returns to $\phi^{\omega*}$.

Next, lets look at preference distribution $\hat{\lambda}$. Condition 2 of proposition 4.9 implies that at the new social norm $\phi^{\omega} < \phi^{\omega*}$, the dis-utility from social disapproval for social norm violation that individuals of preference type $\bar{\theta}$ experience exceeds social disapproval for social norm violation and non-conformity on cultural fitness and, thus, the costs of contribution if only all norm holders cooperate, $\bar{\theta}_s \tilde{v}(\phi^{\omega}) > \gamma\left(v(\phi^{\omega}) + \Delta k(\phi^{\omega})\right) > -\Delta m^{\omega}(\phi^{\omega})$. Cooperation by only all norm holders can no longer be a behavioral equilibrium. Some norm non-holders of preference type $\bar{\theta}$ prefer to cooperate and change their behavioral routine accordingly, $\sigma_{0,\bar{\theta}}^{\omega*} \in (0, 1)$. Since some norm non-holders of preference type $\bar{\theta}$ cooperate and some do not, $\sigma_{0,\bar{\theta}}^{\omega*} \in (0, 1)$, these individuals must be indifferent between both behavioral routines at social norm ϕ^{ω} and the equilibrium cooperation costs, $-\Delta m^{\omega}(\psi^{\omega*}(\phi^{\omega}, \hat{\lambda})) = \bar{\theta}_s \tilde{v}(\phi^{\omega})$. It follows that at social norm ϕ^{ω}, the equilibrium costs of cooperation exceed social disapproval for social norm violation and non-conformity on cultural fitness, $-\Delta m^{\omega}(\psi^{\omega*}(\phi^{\omega}, \hat{\lambda})) = \bar{\theta}_s \tilde{v}(\phi^{\omega}) >$



(a) robustness of a cultural equilibrium of an imperfect social norm (b) instability of an imperfect social norm

Figure 3: biological mutation and cultural equilibrium of an imperfect social norm.

$\gamma\left(v(\phi^\omega)+\Delta k(\phi^\omega)\right)$. Since all norm holders but only some norm non-holders cooperate, this negatively impacts the relative cultural fitness of the norm holders. Moreover, since the social norm is relatively weak, $\phi^\omega < \phi^{\omega*} < \frac{1}{2}$, norm holders are subject to greater social disapproval from non-conformity, $\Delta k(\phi^\omega) < 0$. In conjunction, this implies that the cultural fitness of norm non-holders exceeds that of norm holders, $(\sigma_1^{\omega*} - \sigma_0^{\omega*})(\gamma v(\phi^\omega) + \Delta m^\omega(\psi^\omega(\phi^\omega, \hat{\lambda}))) + \Delta k(\phi^\omega) < 0$. The social norm further weakens at ϕ^ω, implying that $\phi^{\omega*}$ cannot be (part of) a cultural equilibrium at preference distribution $\hat{\lambda}$.

4.3 Discussion

The results of this section demonstrate that the proposed framework can explain the persistence of a diverse culture with varying social norms and cooperation levels across different situations. Moreover, we can utilize the results to investigate how societal characteristics affect cooperation at the cultural level. For simplification purposes, the following discussion focuses on homogeneous approval preferences and the respective cultural equilibria.[14]

Generally speaking, increases in the weight of social approval on cultural fitness γ favor greater levels of cooperation. For example, consider an increase of γ in figure 2. $\gamma(v(\phi^\omega) + \Delta k(\phi^\omega))$ rotates counterclockwise around its x-intercept, which induces it's intersection with $-\Delta m^\omega(\phi^\omega)$ to move to the right.

[14]The discussion can easily be expanded to heterogeneous approval preferences and the results of section 4.2.2.

Similarly, greater social disapproval for social norm violation $v(\phi^\omega)$ also favors the existence of strong social norms. Since individuals who hold the cooperation norm always have more incentives to cooperate, the share of cooperators among them must always be greater than the share among norm non-holders. Therefore, greater social disapproval for norm violation generally impacts cultural fitness differences in favor of the norm holders.

Social disapproval for non-conformity supports the persistence of relatively strong social norms $\phi^{\omega*} > \frac{1}{2}$ even if other forces favor a weakening. By similar reasoning, conformity concerns can trap a relatively weak social norm $\phi^{\omega*} < \frac{1}{2}$, despite other forces supporting a further spread. Consider figure 2 and an increase in the slope of Δk. As a consequence, $\gamma(v(\phi^\omega) + \Delta k(\phi^\omega))$ rotates anticlockwise around it's point at $\frac{1}{2}$, shifting it's intersection with $-\Delta m^\omega(\phi^\omega)$ to the right or left, depending on whether $\phi^{\omega*} > \frac{1}{2}$ or $\phi^{\omega*} < \frac{1}{2}$ respectively. By similar mechanism, social disapproval for non-conformity stabilizes a social norm if it is either perfect or absent.

Social disapproval for hypocrisy h generally hinders the spread of norms. Since only the carriers of the cooperation norm can experience social disapproval from hypocrisy, they have an evolutionary disadvantage. This insight somewhat contrasts with the impact on equilibrium behavior, where greater social concerns for hypocrisy generally favor cooperation. Consequently, an increase in social disapproval from hypocrisy only seems favorable if it increases behavioral incentives without destroying the respective cultural equilibrium.

Lastly, we investigate consequences of the absence of social disapproval for non-conformity and hypocrisy ($\Delta k(x) = 0 \ \forall x \in [0,1]$ and $h = \tilde{h} = 0$). The presented cultural equilibria still exist under the stated conditions if the behavior of both norm populations differs in some neighborhood of the respective social norm ($\sigma_1^{\omega*} \neq \sigma_0^{\omega*}$ for some U of $\phi^{\omega*}$). This always holds for a cultural equilibrium of an imperfect social norm but need not necessarily be the case for cultural equilibrium of perfect or no social norms. If both norm populations behave equally in some neighborhood of the absent or perfect social norm, they obtain the same material payoff and social approval on average. Hence, norm evolution is at rest. All social norms that induce the same behavior in the norm populations form a connected set of rest points. Whether this set is asymptotically stable depends on material costs and social approval at the set's boundaries.

5 Approval Preferences

This section endogenizes the formation of approval preferences. Throughout, we assume that biological evolution occurs significantly slower than cultural evolution, so that norms and behavior always reach equilibria before further changes in the preference distribution occur.

5.1 Evolutionary Framework

Approval preferences evolve through biological reproduction. Like cultural fitness, the fitness that drives biological reproduction is co-determined by material payoff and social approval. Formally, we write an individual's biological fitness as follows.

Definition 5.1 (Biological Fitness)**.** $b_i^\omega(a_i^\omega, n_i^\omega, \psi^\omega, \phi^\omega) = m_i^\omega(a_i^\omega, \psi^\omega) + \rho s_i^\omega(a_i^\omega, n_i^\omega, \phi^\omega)$, where $0 < \rho < \gamma$ is the weight of social approval on biological fitness.[15]

Individuals with relatively high biological fitness have greater access to social and material resources, which positively affects their parenting abilities (Geary et al. 2004; Irons 1979). This increases their reproductive fitness through greater survival chances of their offspring (Buss and Schmitt 1993; Turke 1989; Wiederman 1993) as well as greater chances of finding mating partners (Bereczkei and Csanaky 1996; Shackelford et al. 2005). Similar to cultural evolution, biological evolution follows an imitative dynamics.

Definition 5.2 (Preference Dynamics)**.**

$$\dot{\lambda}_\theta = \lambda_\theta \left(B_\theta(\sigma, \phi) - B_\lambda(\sigma, \phi) \right) \; \forall \theta \in \Theta,$$

where $B_\lambda(\sigma, \phi) = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta B_\theta(\sigma, \phi)$, $B_\theta(\sigma, \phi) = \sum_{\omega \in \Omega} B_\theta^\omega(\sigma^\omega, \phi^\omega)$, $B_\theta^\omega(\sigma^\omega, \phi^\omega) = \phi^\omega B_{1,\theta}^\omega(\sigma_{1,\theta}^\omega, \phi^\omega) + (1 - \phi^\omega) B_{0,\theta}^\omega(\sigma_{0,\theta}^\omega, \phi^\omega)$, and $B_{n,\theta}^\omega(\sigma_{n,\theta}^\omega, \phi^\omega) = \sigma_{n,\theta}^\omega b^\omega(1, n, \psi^\omega, \phi^\omega) + (1 - \sigma_{n,\theta}^\omega) b^\omega(0, n, \psi^\omega, \phi^\omega)$.

Note that $B_\theta(\sigma, \phi)$ and $B_\lambda(\sigma, \phi)$ correspond to the average biological fitness of all individuals with preference type θ and all individuals in society respectively. We define the following in line with the equilibrium notions for behavior and norms.

[15] Appendix A.4 discusses the consequences of relaxing the assumption on ρ.

Definition 5.3 (Biological Equilibrium)**.** A connected, non-empty and closed set Λ of preference distributions λ is a *biological equilibrium* if it is asymptotically stable in 5.2 and there is no $\tilde{\Lambda} \subset \Lambda$ such that $\tilde{\Lambda}$ is asymptotically stable in dynamics 5.2. We may indicate a biological equilibrium by Λ^*.

5.2 Equilibrium Analysis

The following theorem captures the main result of this paper.

Theorem 5.1. *If the set of situations Ω is sufficiently diverse regarding the costs of contribution $\{-\Delta m^\omega\}_{\omega\in\Omega}$, then there exists a set of preference distribution Λ^* and social norms ϕ^* s.t.*

1. *aggregate behavior at all $\lambda \in \Lambda^*$ is as if all individuals were homogeneous regarding approval preferences,*
2. *ϕ^* are cultural equilibria at all $\lambda \in \Lambda$,*
3. *ϕ^* consist of absent, perfect, and imperfect social norms, and*
4. *the dynamic system returns to the social norms ϕ^* and biological equilibrium Λ^* if preference mutation leads to some preference distribution $\hat{\lambda} \notin \Lambda^*$.*

Proof: Follows from proposition B.2 and proposition 5.1.

From the theorem follows that if the set of situations Ω is sufficiently diverse regarding the costs of contribution, there exists a biological equilibrium Λ^* for a diverse culture ϕ^* persists. In the following, we prove theorem 5.1 by example. Hence, we present a biological equilibrium Λ^* and social norms ϕ^* that satisfy the stated conditions. Our procedure is three-folded. Section 5.2.1 presents the preference type that quasi-persists and derives a candidate for a biological equilibrium $\Lambda_p(\phi)$ for any possible social norms ϕ therefrom. Moreover, we discuss how this candidate depends on the social norms ϕ. Section 5.2.2 introduces the social norms ϕ_r^* that shape the potential biological equilibrium in such a way, that it is in fact a biological equilibrium. Section 5.2.3 combines the insights to show that the potential

biological equilibrium of section 5.2.1 indeed corresponds to a biological equilibrium at which the social norms of section 5.2.2 prevail.

5.2.1 The Potential Biological Equilibrium

We start by presenting the preference type that quasi-persists. Based on this preference type and any social norms $\phi \in [0,1]^{|\Omega|}$, we derive a candidate for a biological equilibrium at which the social norms ϕ prevail.

Definition 5.4 (The Dominant Preference Type)**.** $\theta^d := (\rho(1+\delta), \rho(h-(1+\delta)\tilde{h}))$.

We write the preference distribution for which only preference type θ^d exists as λ^d. Any individual with preference type θ^d experiences dis-utility from expressed social disapproval equal to the degree that gossip δ proportionally increases it and impacts biological fitness. Preferences for self-approval account for social disapproval from hypocrisy that arises from information pooling. By substituting θ_s^o and θ_p^o into the utility function, it becomes apparent that the utility of an individual with preference type θ^d mimics biological fitness.

Consider any preference distribution λ for which some individuals have preference type θ^d, $\theta^d \in \text{supp}(\lambda)$. Since equilibrium behavior maximizes an individual's utility, equilibrium behavior of individuals with preference type θ^d maximizes their biological fitness (see lemma B.13 in appendix B.3). Hence, individuals with preference type θ^d always obtain (weakly) greater biological fitness than their peers, $B_{\theta^d}(\sigma^*, \phi) \geq B_\theta(\sigma^*, \phi)\ \forall \theta \in \text{supp}(\lambda)$. Unless all individuals currently maximize their biological fitness, the preference type θ^d spreads in society, $\dot{\lambda}_{\theta^d} = 0 \Leftrightarrow B_{\theta^d}(\sigma^*, \phi) = B_\theta(\sigma^*, \phi)\ \forall \theta \in \text{supp}(\lambda)$. Suppose all individuals behave as if their preference type was θ^d. In that case, all individuals maximize biological fitness, and preference evolution is at rest (see lemma B.14 in appendix B.3). Moreover, if all individuals behave as if their preference type was θ^d at preference distribution λ, then norm population behavior $(\sigma_1^{\omega*}, \sigma_0^{\omega*})$ in any situation ω is as if preferences were distributed according to λ^d. Analogously, if norm population behavior $(\sigma_1^{\omega*}, \sigma_0^{\omega*})$ at preference distribution λ is not as if preferences were distributed according to λ^d, some individuals behave differently from how they would if their preference type was θ^d. They do not maximize biological fitness, implying that their preference type erodes. Given the above argumentation, the following

constitutes an obvious candidate for a potential biological equilibrium at which the social norms ϕ prevail.

Definition 5.5 (Potential Biological Equilibrium)**.** For all $\phi \in [0,1]^{|\Omega|}$, let $\Lambda_p(\phi)$ be the set of preference distributions that satisfy $\lambda \in \Lambda_p(\phi)$ if and only if for all $\omega \in \Omega$,

1. ϕ^ω is a cultural equilibrium at λ and

2. $\sigma_n^\omega = \bar{\sigma}_n^\omega \ \forall n \in \{0,1\} \backslash \{1 - \phi_r^{\omega *}\}, \sigma^\omega \in \Sigma^{\omega *}(\phi^\omega, \lambda), \bar{\sigma}^\omega \in \Sigma^{\omega *}(\phi^\omega, \lambda^d)$.

For any social norms ϕ, the set $\Lambda_p(\phi)$ contains all preference distributions λ so that for each situation $\omega \in \Omega$ (1) the social norm ϕ^ω is a cultural equilibrium and (2) equilibrium norm population behavior is as if only the dominant preference type θ^d existed. Lets look at condition 2 in some more detail. Since at preference distribution λ^d all individuals are homogeneous regarding approval preferences, $\sup(\lambda) = \{\theta^d\}$, the behavioral equilibrium at social norm ϕ^ω is a singleton, $\Sigma^{\omega *}(\phi^\omega, \lambda^d) = \{\bar{\sigma}^{\omega *}\}$ (recall lemma B.7 in appendix B.1). We can rewrite condition 2 of definition 5.5 as follows:

- if $\phi^\omega \in (0,1)$, $(\sigma_1^\omega, \sigma_0^\omega) = (\bar{\sigma}_1^\omega, \bar{\sigma}_0^\omega) \ \forall \sigma^\omega \in \Sigma^{\omega *}(\phi^\omega, \lambda), \bar{\sigma}^\omega \in \Sigma^{\omega *}(\phi^\omega, \lambda^d)$ and

- if $\phi^\omega \in \{0,1\}$, $\psi^{\omega *}(1, \lambda) = \psi^{\omega *}(1, \lambda^d)$.

First, if both personal norms exist, $\phi^\omega \in (0,1)$, then norm population behavior $(\sigma_1^\omega, \sigma_0^\omega)$ of any Nash equilibrium σ^ω at preference distribution λ is the same as the equilibrium norm population behavior at λ^d, $(\sigma_1^\omega, \sigma_0^\omega) = (\bar{\sigma}_1^\omega, \bar{\sigma}_0^\omega)$. Second, if all individuals share the same personal norm, $n_i = \phi^\omega \in \{0,1\} \, \forall i \in \mathcal{I}$, actual norm population behavior is fully described by only the existing norm population's behavior $\sigma_{\phi^\omega}^{\omega *}$. Moreover, the behavior of the existing norm population fully defines the total equilibrium cooperation share, $\sigma_{\phi^\omega}^{\omega *} = \psi^{\omega *}(\phi^\omega, \lambda)$.

Note how the social norms ϕ play a major role in shaping the potential biological equilibrium $\Lambda_p(\phi)$. For each situation ω, the social norm ϕ^ω and cost function m^ω co-determine equilibrium behavior $\sigma^{\omega *}$ at preference distribution λ^d. In turn, equilibrium behavior at λ^d defines which preference distribution λ is an element of $\Lambda_p(\phi)$. Below, we discuss some consequences regarding the $\Lambda_p(\phi)$ structure following form the social norms ϕ.

Lemma 5.1. *Consider any* $\phi \in [0,1]^{|\Omega|}$. $\lambda \in \Lambda_p(\phi)$ *implies that for all* $\theta \in \text{supp}(\lambda)$ *and* $\omega \in \Omega$ *the following holds:*

1. $\theta_s \leq \min\{\frac{-\Delta m^\omega(\phi^\omega)}{\tilde{v}(\phi^\omega)} : \phi^\omega = \psi^{\omega*}(\phi^\omega, \lambda^d) \in (0,1)\}$, *and*
2. $\theta_s \tilde{v}(\phi^\omega) + \theta_s \tilde{h} + \theta_p \geq (\tilde{v}(\phi^\omega) + \tilde{h}) \frac{\max_{\omega \in \Omega}\{-\Delta m^\omega(1) : \phi^\omega = \psi^{\omega*}(1,\lambda^d)=1\}}{\tilde{v}(1)+\tilde{h}}$.

Proof: see appendix B.3.

Lemma 5.1 states two necessary conditions that must hold for each preference type θ at any preference distribution $\lambda \in \Lambda_p(\phi)$. In particular, it presents (1) an upper bound on preferences for social approval for all individuals and (2) a lower bound on utility benefits from cooperating for all norm holders in any situation ω. Below, we discuss these conditions in more detail.

The first condition describes an upper bound on preferences for social approval that arises from the existence of a cultural equilibrium of an imperfect social norm $\phi^\omega \in (0,1)$ for which all individuals at preference distribution λ^d cooperate if and only if they hold the cooperation norm, $(\sigma_1^{\omega*}, \sigma_0^{\omega*}) = (1,0) \Rightarrow \psi^{\omega*}(\phi^\omega, \lambda^d) = \phi^\omega$. For any preference distribution λ to mimic equilibrium behavior at λ^d in situation ω, all individuals must (weakly) prefer to cooperate if they hold the cooperation norm and (weakly) prefer to defect if they do not hold the cooperation norm at λ. Formally, this is true if all existing preference types $\theta \in \text{supp}(\lambda)$ yield utility benefits from cooperating that are weakly smaller than the costs of cooperation for norm non-holders and weakly greater than the costs of cooperation for norm holders, $\theta_s \tilde{v}(\phi^\omega) \leq -\Delta m^\omega(\phi^\omega) \leq \theta_s \tilde{v}(\phi^\omega) + \theta_s \tilde{h} + \theta_p \ \forall \theta \in \text{supp}(\lambda)$. Formally, we can rearrange the left part of the inequality to state that preferences for social approval must be sufficiently small for any preference type $\theta \in \text{supp}(\lambda)$, $\theta_s \leq \frac{-\Delta m^\omega(\phi^\omega)}{\tilde{v}(\phi^\omega)} \ \forall \theta \in \text{supp}(\lambda)$. Condition 1 then follows from accounting for all such situations.

The second condition describes a lower bound on utility benefits from cooperating for norm holders that arises from the existence of a perfect social norm $\phi^{\hat{\omega}} = 1$ inducing full cooperation at preference distribution λ^d, $\psi^{\hat{\omega}*}(1, \lambda^d) = 1$. Since the social norm is perfect, everyone holds the cooperation norm for situation $\hat{\omega}$. For equilibrium behavior at λ to mimic that of λ^d, the respective equilibrium cooperation share must be one too, $\psi^{\hat{\omega}*}(1, \lambda) = 1$. This

is true if all existing preference types $\theta \in \text{supp}(\lambda)$ induce norm-based utility benefits for norm holders that outweigh material costs at full cooperation, $\theta_p + \theta_s\tilde{h} + \theta_s\tilde{v}(1) \geq -\Delta m^{\hat{\omega}}(1)$. This condition holds for all such situations $\hat{\omega}$, if it holds for the most costly one among them, $\theta_p + \theta_s\tilde{h} + \theta_s\tilde{v}(1) \geq \max_{\omega\in\Omega}\{-\Delta m^{\omega}(1) : \phi^{\omega} = \psi^{\omega *}(1, \lambda^d) = 1\}$. Hence, we obtain a lower bound on norm-based utility benefits at the perfect social norm for an individual holding the cooperation norm. We can rearrange the condition for any preference type $\theta \in \text{supp}(\lambda)$ to state that given the type's preferences for self-approval θ_p, preferences for social approval must be sufficiently large, $\theta_s \geq \frac{\max_{\omega\in\Omega}\{-\Delta m^{\omega}(1):\phi^{\omega}=\psi^{\omega *}(1,\lambda^d)=1\}-\theta_p}{\tilde{v}(1)+\tilde{h}}$. Consider any other situation ω with social norm ϕ^{ω}. From the above, we can derive that in that situation ω, an individual with preferences type $\theta \in \text{supp}(\lambda)$ who holds the cooperation norm obtains utility benefits from cooperating that satisfy $\theta_s\tilde{v}(\phi^{\omega}) + \theta_s\tilde{h} + \theta_p \geq \frac{\tilde{v}(\phi^{\omega})+\tilde{h}}{\tilde{v}(1)+\tilde{h}}(\max_{\omega\in\Omega}\{-\Delta m^{\omega}(1) : \phi^{\omega} = \psi^{\omega *}(1, \lambda^d) = 1\} - \theta_p) + \theta_p$. Since $\frac{\tilde{v}(\phi^{\omega})+\tilde{h}}{\tilde{v}(1)+\tilde{h}} \leq 1$, the right side of the inequality is at a minimum for $\theta_p = 0$. Hence, we get that in any situation ω, the utility benefits from cooperating for norm holders of any preference type $\theta \in \text{supp}(\lambda)$ are bounded below by $\frac{\tilde{v}(\phi^{\omega})+\tilde{h}}{\tilde{v}(1)+\tilde{h}} \max_{\omega\in\Omega}\{-\Delta m^{\omega}(1) : \phi^{\omega} = \psi^{\omega *}(1, \lambda^d) = 1\}$. This coincides with condition 2 of lemma 5.1.

5.2.2 Stability-Inducing Equilibrium Culture

The previous section presented the potential biological equilibrium and illustrated the social norms' importance in shaping it. This section presents the social norms ϕ_r^* that shape $\Lambda_p(\phi_r^*)$ in such a way that $\Lambda_p(\phi_r^*)$ is indeed a biological equilibrium. We here focus on introducing the social norms ϕ_r^* and discussing how they shape the potential biological equilibrium $\Lambda_p(\phi_r^*)$.

Definition 5.6 (Stability-Inducing Equilibrium Culture)**.** Let ϕ_r^* be such that

1. for all $\omega \in \Omega$, $\phi_r^{\omega *}$ is a cultural equilibrium of proposition 4.1, proposition 4.2, or proposition 4.3 at λ^d.

2. for $\bar{\omega} := \text{argmin}_{\omega\in\Omega}\{\frac{-\Delta m^{\omega}(\phi_r^{\omega *})}{\tilde{v}(\phi_r^{\omega *})} : \phi_r^{\omega *} \in (0, 1)\}$,

   (a) $\phi_r^{\bar{\omega} *} < \frac{1}{2}$ and

   (b) $\frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega} *})}{\tilde{v}(\phi_r^{\bar{\omega} *})} \times \frac{\mathrm{d}\tilde{v}(x)}{\mathrm{d}x}|_{x=\phi_r^{\omega *}} < \gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi_r^{\omega *}} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi_r^{\omega *}})$,

3. $\frac{\max_{\omega\in\Omega}\{-\Delta m^{\omega}(1):\psi^{\omega*}=1\}}{\tilde{h}+\tilde{v}(1)}(\tilde{h}+\tilde{v}(\phi_r^{\omega*})) > -\Delta m^{\omega}(\phi_r^{\omega*})\ \forall \phi_r^{\omega*} \in (0,1)$, and

4. (a) $\min_{\omega\in\Omega}\{\psi^{\omega*}(1,\lambda^d) : \psi^{\omega*} \leq \phi_r^{\omega*} = 1\} > \frac{h-\Delta k(1)}{h}$ or

   (b) $\min_{\omega\in\Omega}\{\frac{-\Delta m^{\omega}(\phi_r^{\omega*})}{\tilde{v}(\phi_r^{\omega*})} : \phi_r^{\omega*} \in (0,1)\}\tilde{v}(1) \leq \theta_s^d\tilde{v}(1) + \theta_s^d\tilde{h} + \theta_p^d$.

The first condition states that each social norm $\phi_r^{\omega*}$ is a cultural equilibrium of section 4.2.1 at the homogeneous preference distribution λ^d. This implies that $\Lambda_p(\phi_r^*)$ is non-empty, since λ^d is always in it. The second condition states that proposition 4.9 applies to the situation $\bar{\omega}$ that introduces the lowest upper bound on preferences for social approval θ_s for all preference types θ that may occur in the potential biological equilibrium $\Lambda_p(\phi_r^*)$ (recall lemma 5.1). Hence, if in situation $\bar{\omega}$ and at social norm $\phi_r^{\bar{\omega}*}$ all individuals cooperate if and only if they hold the cooperation norm, but some norm non-holders are indifferent between both behavioral routines, then the social norm $\phi_r^{\bar{\omega}*}$ is not a cultural equilibrium. Condition 3 states that the lower bound on utility benefits from cooperating for norm holders of any preference type θ that may occur in the potential biological equilibrium $\Lambda_p(\phi_r^*)$ (recall lemma 5.1) exceeds the costs of cooperation for all situations with a cultural equilibrium of an imperfect social norm. Condition 4 is two-folded. Condition 4a considers the situation ω that induces the lowest equilibrium cooperation share $\psi^{\bar{\omega}*}(1,\lambda^d)$ at preference distribution λ^d among all situations with a perfect social norm $\phi_r^{\omega*} = 1$. It states that social disapproval for non-conformity outweighs average social disapproval for hypocrisy, $\Delta k(1) > (1-\psi^{\bar{\omega}*}(1,\lambda^d))h$. Condition 4b states that the upper bound on social approval preferences θ_s for all preference types θ that may occur in the potential biological equilibrium $\Lambda_p(\phi_r^*)$ (recall lemma 5.1) is so small that at the perfect social norm all norm-non holders of such a preference type θ would obtain fewer utility benefits from cooperating than norm holders of the dominant preference type θ^d. Proposition B.2 in appendix B.3 includes a proof that if the set of situations is sufficiently large, then there exists a ϕ_r^* of definition 5.6 that features all three types of cultural equilibria of section 4.2.1 at preference distribution λ^d.

From lemma 5.1, we can derive some insights on how the social norms ϕ_r^* shape $\Lambda_p(\phi_r^*)$.

Lemma 5.2. *Consider any $\Lambda_p(\phi_r^*)$ and ϕ_r^* satisfying definition 5.5 and 5.6 respectively. For all $\omega \in \Omega$, $\lambda \in \Lambda_p(\phi_r^*)$, and $\theta \in \text{supp}(\lambda)$:*

- *$\phi_r^{\omega*} \in (0,1)$ implies that $\theta_s\tilde{v}(\phi_r^{\omega*}) < -\Delta m^{\omega}(\phi_r^{\omega*}) < \theta_s\tilde{v}(\phi_r^{\omega*}) + \theta_s\tilde{h} + \theta_p$.*

- $\phi_r^{\omega*} = 1$ *implies that*

  1. $((1 - \psi^{\omega*}(1, \lambda))h < \Delta k(1)$ *or*
  2. $\theta_s \tilde{v}(1) < -\Delta m^{\omega}(\psi^{\omega*}(1, \lambda))$.

Proof: see appendix B.3.

The first condition states that at any preference distribution $\lambda \in \Lambda_p(\phi_r^*)$ and situation ω with an imperfect social norm $\phi_r^{\omega*} \in (0, 1)$, all norm non-holders strictly prefer to defect and all norm holders strictly prefer to cooperate. The second condition states that at any preference distribution $\lambda \in \Lambda_p(\phi_r^*)$ and situation ω with a perfect social norm $\phi_r^{\omega*} = 1$, either (1) social disapproval for non-conformity outweighs average social disapproval for hypocrisy or (2) the costs of contribution outweigh social disapproval for social norm violation on utility for all preference types $\theta \in \text{supp}(\lambda)$. Below, we look at both conditions in more detail and argue why they hold. We consider any preference distribution $\lambda \in \Lambda_p(\phi_r^*)$ to do so.

First, we consider condition 1. Recall from lemma 5.1 that in any situation ω there is a lower bound on utility benefits from cooperating for norm holders of any preference type $\theta \in \text{supp}(\lambda)$. The third condition of definition 5.6 requires that this lower bound exceeds the costs of cooperation in all situations ω with an imperfect social norm $\phi_r^{\omega*} \in (0, 1)$. Hence, it must hold that norm holders of any preference type $\theta \in \text{supp}(\lambda)$ strictly prefer to cooperate at any imperfect social norm $\phi_r^{\omega*} \in (0, 1)$, $-\Delta m^{\omega}(\phi_r^{\omega*}) < \theta_s \tilde{v}(\phi_r^{\omega*}) + \theta_s \tilde{h} + \theta_p$. Hence, the right inequality of condition 1 of lemma 5.2 is true. Next, recall from lemma 5.1 that a situation ω with an imperfect social norm $\phi_r^{\omega*} \in (0, 1)$ bounds preferences for social approval above for all preference types $\theta \in \text{supp}(\lambda)$, $\theta_s \leq \frac{-\Delta m^{\omega}(\phi_r^{\omega*})}{\tilde{v}(\phi_r^{\omega*})}$. Consider some situation ω with an imperfect social norm $\phi_r^{\omega*} \in (0, 1)$ that does not introduce the lowest upper bound on social approval preferences, $\frac{-\Delta m^{\omega}(\phi_r^{\omega*})}{\tilde{v}(\phi_r^{\omega*})} > \min_{\bar{\omega} \in \Omega}\{\frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*})}{\tilde{v}(\phi_r^{\bar{\omega}*})} : \phi_r^{\bar{\omega}*} \in (0, 1)\}$. Hence, there exists some other situation $\bar{\omega}$ with a cultural equilibrium of an imperfect social norm that induces smaller social approval preferences, $\theta_s \leq \min_{\bar{\omega} \in \Omega}\{\frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*})}{\tilde{v}(\phi_r^{\bar{\omega}*})} : \phi_r^{\bar{\omega}*} \in (0, 1)\} < \frac{-\Delta m^{\omega}(\phi_r^{\omega*})}{\tilde{v}(\phi_r^{\omega*})}$. It follows that in situation ω, the costs of contribution strictly outweigh utility benefits from cooperating for norm non-holders, $\theta_s \tilde{v}(\phi_r^{\omega*}) < -\Delta m^{\omega}(\phi_r^{\omega*})$. Alternatively, consider the situation ω that introduces the lowest upper bound on preferences for social approval,

$\frac{-\Delta m^{\omega}(\phi_r^{\omega*})}{\tilde{v}(\phi_r^{\omega*})} = \min_{\bar{\omega}\in\Omega}\{\frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*})}{\tilde{v}(\phi_r^{\bar{\omega}*})} : \phi_r^{\bar{\omega}*} \in (0,1)\}$. Consequently, the second condition of definition 5.6 applies to that situation ω. From the above, we know that for all preference types $\theta \in \text{supp}(\lambda)$ the following holds: $\theta_s\tilde{v}(\phi_r^{\omega*}) \leq -\Delta m^{\omega}(\phi_r^{\omega*}) < \theta_s\tilde{v}(\phi_r^{\omega*}) + \theta_s\tilde{h} + \theta_p$. In conjunction with condition 2 of definition 5.6, proposition 4.9 implies that $\phi_r^{\omega*}$ is not a cultural equilibrium if some norm non-holders are indifferent between cooperating and defecting at social norm $\phi_r^{\omega*}$. By definition 5.5 of $\Lambda_p(\phi_r^*)$, a preference distribution for which $\phi_r^{\omega*}$ is not a cultural equilibrium is not an element of the potential biological equilibrium $\Lambda_p(\phi_r^*)$. Hence, for all $\lambda \in \Lambda_p(\phi_r^*)$ it must hold that social approval preferences of all existing preference types $\theta \in \text{supp}(\lambda)$ satisfy $\theta_s < \frac{-\Delta m^{\omega}(\phi^{\omega*})}{\tilde{v}(\phi^{\omega*})}$. Rearranging then yields the left inequality of condition 1 of lemma 5.2.

Next, we look at condition 2 of lemma 5.2. This condition follows from the fourth condition on $\phi_r^{\omega*}$ in definition 5.6. Consider any situation ω with a cultural equilibrium of a perfect social norm $\phi_r^{\omega*} = 1$. Since λ is an element of the potential biological equilibrium $\Lambda_p(\phi_r^*)$, the cooperation share at the perfect social norm $\phi_r^{\omega*} = 1$ for preference distributions λ and λ^d equal, $\psi^{\omega*}(1,\lambda) = \psi^{\omega*}(1,\lambda^d)$. If condition 4a of definition 5.6 holds, then the cooperation share in any situation ω with a perfect social norm $\phi_r^{\omega*} = 1$ satisfies $\psi^{\omega*}(1,\lambda) > \frac{\Delta k(1)-h}{h}$. Rearranging yields that average social disapproval for hypocrisy is smaller than social disapproval for non-conformity, $((1-\psi^{\omega*}(1,\lambda))h < \Delta k(1)$. Condition 2 of lemma 5.2 is true. Alternatively, suppose that condition 4a of definition 5.6 does not hold, but condition 4b does. Condition 4a does only not hold if cooperation is partial, $\psi^{\omega*}(1,\lambda) = \psi^{\omega*}(1,\lambda^d) < 1$. Hence, some individuals cooperate and some defect at preference distribution λ^d and the perfect social norm $\phi_r^{\omega*} = 1$. Consequently, individuals of preference type θ^d are indifferent between both behavioral routines when holding the cooperation norm, $-\Delta m^{\omega}(\psi^{\omega*}(1,\lambda^d)) = \theta_s^d\tilde{v}(1) + \theta_s^d\tilde{h} + \theta_p^d$. From condition 1 of lemma 5.2 follows that preferences for social approval of any preference type $\theta \in \text{supp}(\lambda)$ are bounded above, $\theta_s < \min_{\omega\in\Omega}\{\frac{-\Delta m^{\omega}(\phi_r^{\omega*})}{\tilde{v}(\phi_r^{\omega*})} : \phi_r^{\omega*} \in (0,1)\}$. Consequently, condition 4b of definition 5.6 implies that this upper bound yields $\theta_s\tilde{v}(1) < \min_{\omega\in\Omega}\{\frac{-\Delta m^{\omega}(\phi_r^{\omega*})}{\tilde{v}(\phi_r^{\omega*})} : \phi_r^{\omega*} \in (0,1)\}\tilde{v}(1) \leq \theta_s^d\tilde{v}(1) + \theta_s^d\tilde{h} + \theta_p^d = -\Delta m^{\omega}(\psi^{\omega*}(1,\lambda^d))$. Thus, condition 2 of lemma 5.2 is true.

Lemma 5.3. *For all $\delta > 0$ there is some U of $\Lambda_p(\phi_r^*)$ s.t. for all $\omega \in \Omega$ and $\hat{\lambda} \in U$:*

1. *If* $\phi_r^{\omega*} \in \{0,1\}$*, then* $\phi_r^{\omega*}$ *is a cultural equilibrium at* $\hat{\lambda}$*.*

2. *If* $\phi_r^{\omega*} \in (0,1)$*, then there is a cultural equilibrium* $\hat{\Phi}^{\omega*}$ *at* $\hat{\lambda}$ *s.t. for all* $\hat{\phi}^{\omega} \in \hat{\Phi}^{\omega*}$*,* $|\phi_r^{\omega*} - \hat{\phi}^{\omega}| < \delta$*.*

Proof: The lemma is a corollary of proposition 4.4, proposition 4.6, proposition 4.8, and lemma 5.2.

Finally, lemma 5.3 presents some consequences that follow from lemma 5.2 in conjunction with the results of section 4.2.2. Lemma 5.3 states that for any preference distribution $\hat{\lambda}$ close to $\Lambda_p(\phi_r^*)$, there exist some cultural equilibria $\hat{\phi}^*$ that are close to ϕ_r^*. In particular, for any situation ω with a cultural equilibrium of no or a perfect social norm $\phi_r^* \in \{0,1\}$, the respective social norm ϕ_r^* is also a cultural equilibrium at $\hat{\lambda}$. This follows directly from proposition 4.4 and proposition 4.6. For any situation ω with an imperfect social norm $\phi_r^* \in (0,1)$, there exists a cultural equilibrium $\hat{\Phi}^{\omega*}$ at $\hat{\lambda}$ that is close to ϕ_r^*. This follows from proposition 4.8.

5.2.3 The Biological Equilibrium and Prevailing Social Norms

This section proceeds with the evolutionary analysis of $\Lambda_p(\phi_r^*)$. In particular, we combine the previous insights to show that $\Lambda_p(\phi_r^*)$ is a biological equilibrium at which the social norms ϕ_r^* prevail. Formally, we will show that the following proposition is true.

Proposition 5.1 (Biological Equilibrium and prevailing Social Norms)**.** *Consider any* ϕ_r^* *of definition 5.6. All triplets* $(\sigma, \phi_r^*, \lambda)$ *s.t. (1)* $\sigma^{\omega} \in \Sigma^{\omega*}(\phi_r^*, \lambda^d)\ \forall \omega \in \Omega$ *and (2)* $\lambda \in \Lambda_p(\phi_r^*)$ *form an asymptotically stable set of rest points.*

Proof: see appendix B.3.

As a starting point, suppose the social norms correspond to ϕ_r^* satisfying definition 5.6, preferences are distributed according to $\lambda \in \Lambda_p(\phi_r^*)$, and all individuals behave as if their preference type was θ^d. Since the respective social norms and behavior are in equilibrium at $\lambda \in \Lambda_p(\phi_r^*)$, their evolutionary processes are at rest. Moreover, everyone maximizes their biological fitness since all individuals behave as if their preference type was θ^d. Biological

evolution is also at rest. Hence, the dynamic system as a whole is at rest. Random walk of preferences across the set $\Lambda_p(\phi_r^*)$ neither alters equilibrium norm population behavior nor the cultural equilibria. All individuals keep maximizing their biological fitness. The dynamic system remains at rest. $\Lambda_p(\phi_r^*)$ constitutes a set of rest points where the social norms ϕ_r^* prevail.

We must investigate what happens when random mutation yields some preference distribution outside of $\Lambda_p(\phi_r^*)$. Consider the appearance of some biological mutants that leads to a change from preference distribution $\lambda \in \Lambda_p(\phi_r^*)$ to $\hat{\lambda} \notin \Lambda_p(\phi_r^*)$. At the post-mutation preference distribution $\hat{\lambda} \notin \Lambda_p(\phi_r^*)$, the social norms ϕ_r^* may no longer constitute cultural equilibria. Consequently, culture evolves to the (possibly new) cultural equilibrium $\hat{\Phi}^{\omega*}$ in each situation ω. Similarly, behavior coordinates into a new behavioral equilibrium $\Sigma^{\omega*}(\hat{\phi}^{\omega}, \hat{\lambda})$ for each social norm $\hat{\phi}^{\omega} \in \hat{\Phi}^{\omega*}$ and preference distribution $\hat{\lambda}$. The set $\Lambda_p(\phi_r^*)$ is a biological equilibrium at which the social norms ϕ_r^* prevail, if

1. at any post-mutation social norms $\hat{\phi} \in \prod_{\omega\in\Omega} \hat{\Phi}^{\omega*}$ and behavior $\hat{\sigma} \in \prod_{\omega\in\Omega} \Sigma^{\omega*}(\hat{\phi}^{\omega}, \hat{\lambda})$[16], biological evolution drives preferences back towards some element in the potential biological equilibrium and

2. the social norms return to ϕ_r^* once preferences return to $\Lambda_p(\phi_r^*)$.

If the social norms and preference distribution return to ϕ_r^* and $\Lambda_p(\phi_r^*)$, then equilibrium behavior inevitably returns to $\Sigma^{\omega*}(\phi_r^*, \lambda^d)$ in each situation ω. This holds since $\Sigma^{\omega*}(\phi_r^*, \lambda^d)$ is the unique behavioral equilibrium. Hence, the two conditions stated above imply that the dynamic system as a whole returns to some state $(\sigma, \phi_r^{\omega*}, \lambda)$ for which (1) $\sigma^{\omega} \in \Sigma^{\omega*}(\phi_r^{\omega*}, \lambda^d)$ $\forall\omega \in \Omega$ and (2) $\lambda \in \Lambda_p(\phi_r^*)$.

We continue by showing that biological evolution drives preferences back to some element in $\Lambda_p(\phi_r^*)$ at any post-mutation norms $\hat{\phi}$ and behavior $\hat{\sigma}$. Intuitively, biological evolution yields a return to $\Lambda_p(\phi_r^*)$ if the biological mutants obtain strictly less biological fitness than their peers on average. Formally, this corresponds to $B_{\lambda}(\hat{\sigma}, \hat{\phi}) > B_{\hat{\lambda}}(\hat{\sigma}, \hat{\phi})$ (see Weibull 1997 among others). Below, we inspect biological fitness differences in the different situations

[16] $\prod_{\omega\in\Omega} \hat{\Phi}^{\omega*}$ and $\prod_{\omega\in\Omega} \Sigma^{\omega*}(\hat{\phi}^{\omega}, \hat{\lambda})$ indicate the Cartesian products of all cultural equilibria $\hat{\Phi}^{\omega*}$ and behavioral equilibria $\Sigma^{\omega*}(\hat{\phi}^{\omega}, \hat{\lambda})$ respectively.

$\omega \in \Omega$.

Lemma 5.4. *Consider any ϕ_r^* of definition 5.6 and $\Lambda_p(\phi_r^*)$ of definition 5.5. For each preference distribution $\hat{\lambda} \in [0,1]^{|\Theta|}$ and situation $\omega \in \Omega$, let $\hat{\Phi}^{\omega*}$ be the cultural equilibrium at $\hat{\lambda}$ s.t. $\phi_r^{\omega*}$ is in it's basin of attraction.*

For all $\lambda \in \Lambda_p(\phi_r^*)$, there is some neighborhood U of λ s.t. for all $\hat{\lambda} \in U$ and $\omega \in \Omega$: $B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) = B_{\hat{\lambda}}^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega)$ for all $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ and $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$ if

1. *$\hat{\Phi}^{\omega*} = \{\phi_r^{\omega*}\}$ and*
2. *$\hat{\sigma}_n^\omega = \sigma_n^\omega \ \forall n \in \{0,1\}\backslash\{1-\phi_r^{\omega*}\}, \hat{\sigma}^\omega \in \Sigma^\omega(\hat{\phi}^{\omega*}, \hat{\lambda}), \sigma^\omega \in \Sigma^\omega(\phi^{\omega*}, \lambda^d)$.*

Proof: see appendix B.3.

Lemma 5.4 investigates all situations ω for which after the biological mutation to $\hat{\lambda}$, (1) the social norm $\phi_r^{\omega*}$ is still a cultural equilibrium and (2) all individuals behave as if their preference type was the dominant one at social norm $\phi_r^{\omega*}$ and preference distribution $\hat{\lambda}$. In such a situation, the biological mutation to $\hat{\lambda}$ does neither alter the social norm, $\hat{\Phi}^{\omega*} = \{\phi_r^{\omega*}\}$, nor equilibrium norm population behavior of the existing norm populations, $\hat{\sigma}_n^\omega = \sigma_n^\omega \ \forall n \in \{0,1\}\backslash\{1-\phi_r^{\omega*}\}, \hat{\sigma}^\omega \in \Sigma^\omega(\hat{\phi}^\omega, \hat{\lambda}), \sigma^\omega \in \Sigma^\omega(\phi^\omega, \lambda^d)$. All individuals still maximize biological fitness in that situation, implying that it creates no differences in biological fitness for mutants and non-mutants. Formally, the average biological fitness of preference distributions λ and $\hat{\lambda}$ equals, $B_\lambda^\omega(\hat{\phi}^\omega, \hat{\lambda}) = B_{\hat{\lambda}}^\omega(\hat{\phi}^\omega, \hat{\lambda})$.

We must investigate situations ω for which biological mutation alters (1) the social norm or (2) equilibrium norm population behavior. Note that such a situation must exist, since otherwise $\hat{\lambda}$ would be an element of $\Lambda_p(\phi_r^*)$ (by definition 5.5). Moreover, such a situation ω cannot refer to a situation with an absent social norm $\phi_r^{\omega*} = 0$, since the absent social norm is always a cultural equilibrium, $\phi^{\omega*} = 0 \Rightarrow \hat{\Phi}^{\omega*} = \{0\}$, that induces equilibrium behavior of full defection, $\psi^{\omega*}(0, \lambda) = 0$.

Lemma 5.5. *Consider any ϕ_r^* of definition 5.6 and $\Lambda_p(\phi_r^*)$ of definition 5.5. For each preference distribution $\hat{\lambda} \in [0,1]^{|\Theta|}$ and situation $\omega \in \Omega$, let $\hat{\Phi}^{\omega*}$ be the cultural equilibrium at $\hat{\lambda}$ s.t. $\phi_r^{\omega*}$ is in it's basin of attraction.*

For all $\lambda \in \Lambda_p(\phi_r^*)$, there is some neighborhood U of λ s.t. for all $\hat{\lambda} \in U$ and $\omega \in \{x \in \Omega : \phi_r^{x*} = 1\}$: $B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) > B_{\hat{\lambda}}^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega)$ for all $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ and $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$ if

1. *$\hat{\Phi}^{\omega*} \neq \{\phi_r^{\omega*}\}$ or*
2. *$(\hat{\sigma}_1^\omega, \hat{\sigma}_0^\omega) \neq (\sigma_1^\omega, \sigma_0^\omega)$ for some $n \in \{0,1\} \backslash \{1 - \phi_r^{\omega*}\}, \hat{\sigma}^\omega \in \Sigma^\omega(\hat{\phi}^{\omega*}, \hat{\lambda}), \sigma^\omega \in \Sigma^\omega(\phi^{\omega*}, \lambda^d)$.*

Proof: see appendix B.3.

Lemma 5.5 considers all situations ω with a cultural equilibrium of a perfect social norm $\phi_r^{\omega*} = 1$ at $\Lambda_p(\phi_r^*)$. Since the post-mutation preference distribution $\hat{\lambda}$ is close to λ, the perfect social norm is also a cultural equilibrium at $\hat{\lambda}$ (recall lemma 5.3). Biological mutation leaves the social norm unaltered, $\hat{\Phi}^{\omega*} = \{1\}$, so that all individuals hold the cooperation norm. Hence, condition 1 of the lemma never holds, $\hat{\phi}^\omega = \phi_r^{\omega*} = 1 \, \forall \hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$. Suppose condition 2 holds. Since the social norm is perfect, $\hat{\phi}$ the condition implies that the cooperation level at preference distribution $\hat{\lambda}$ must differ from that at λ, $\psi^{\omega*}(1, \lambda) \neq \psi^{\omega*}(1, \hat{\lambda})$. Moreover, note that to analyze differences in biological fitness, we only need to investigate norm holders' biological fitness and behavior.

First, suppose all individuals cooperate before the biological mutation occurs, $\psi^{\omega*}(1, \lambda) = 1$. Since all individuals maximize their biological fitness at preference distribution λ, cooperating yields (weakly) greater biological fitness than defecting, $b^\omega(1,1,1,1) \geq b^\omega(0,1,1,1)$. The biological mutation to preference distribution $\hat{\lambda}$ yields some of the biological mutants to prefer defecting rather than cooperating. hence, the cooperation share decreases, $\psi^{\omega*}(1, \hat{\lambda}) < 1$. The decrease in the cooperation share reduces the costs of cooperation, $-\Delta m^\omega(\psi^{\omega*}(1, \hat{\lambda})) < -\Delta m^\omega(1)$, while the social norm and, thus, social disapproval from defecting remain unchanged. This implies that cooperating yields strictly greater biological fitness than defecting after the biological mutation, $b^\omega(1, 1, \psi^{\omega*}(1, \hat{\lambda}), 1) > b^\omega(0, 1, \psi^{\omega*}(1, \hat{\lambda}), 1)$. Consequently, the defecting biological mutants obtain less biological fitness than their peers in situation ω. It follows that average biological fitness of preference distribution λ outweighs that of $\hat{\lambda}$ at the post-mutation norm and equilibrium behavior in situation ω, $B_\lambda^\omega(\hat{\sigma}^\omega, 1) > B_{\hat{\lambda}}^\omega(\hat{\sigma}^\omega, 1) \, \forall \hat{\sigma}^\omega \in \Sigma^{\omega*}(1, \hat{\lambda})$.

Next, suppose cooperation at preference distribution λ is partial, $\psi^{\omega*}(1,\lambda) < 1$. Since all individuals maximize biological fitness at preference distribution λ, both behavioral routines yield the same biological fitness, $b(1,1,\psi^{\omega*}(1,\lambda),1) = b(0,1,\psi^{\omega*}(1,\lambda),1)$. Suppose biological mutation to preference distribution $\hat{\lambda}$ increases the cooperation level, $\psi^{\omega*}(1,\hat{\lambda}) > \psi^{\omega*}(1,\lambda)$. Hence, some biological mutants who previously defected now cooperate. The costs of contribution increase, $-\Delta m^{\omega}(\psi^{\omega*}(1,\hat{\lambda})) > -\Delta m^{\omega}(\psi^{\omega*}(1,\lambda))$. Since social disapproval is unchanged but the costs of cooperation increase, defecting becomes superior in terms of biological fitness, $b(1,1,\psi^{\omega*}(1,\hat{\lambda}),1) < b(0,1,\psi^{\omega*}(1,\hat{\lambda}),1)$. The share of cooperators among the biological mutants must be larger than that of the non-mutants since the biological mutation increased the overall cooperation level. Hence, the biological mutants obtain less biological fitness than their peers on average. It follows that average biological fitness of preference distribution λ outweighs that of $\hat{\lambda}$ at the post-mutation norm and equilibrium behavior in situation ω, $B^{\omega}_{\lambda}(\hat{\sigma}^{\omega},1) > B^{\omega}_{\hat{\lambda}}(\hat{\sigma}^{\omega},1)\,\forall \hat{\sigma}^{\omega} \in \Sigma^{\omega*}(1,\hat{\lambda})$. The case of a decrease in the cooperation level works analogously.

We continue by looking at all situations ω with an imperfect social norm $\phi_r^{\omega*} \in (0,1)$ at $\Lambda_p(\phi_r^*)$. To do so, we first establish sufficient conditions for the biological mutants to obtain less biological fitness than their peers.

Lemma 5.6. *Consider any ϕ_r^* of definition 5.6 and $\Lambda_p(\phi_r^*)$ of definition 5.5. For each preference distribution $\hat{\lambda} \in [0,1]^{|\Theta|}$, let $\hat{\Phi}^{\omega*}$ be the cultural equilibrium at $\hat{\lambda}$ s.t. ϕ_r^* is in it's basin of attraction.*

For all $\lambda \in \Lambda_p(\phi_r^*)$, there is some neighborhood U of λ s.t. for all $\hat{\lambda} \in U$ and $\omega \in \{x \in \Omega : \phi_r^{x*} \in (0,1)\}$:
$B^{\omega}_{\lambda}(\hat{\sigma}^{\omega},\hat{\phi}^{\omega}) > B^{\omega}_{\hat{\lambda}}(\hat{\sigma}^{\omega},\hat{\phi}^{\omega})$ for all $\hat{\phi}^{\omega} \in \hat{\Phi}^{\omega*}$ and $\hat{\sigma}^{\omega} \in \Sigma^{\omega*}(\hat{\phi}^{\omega},\hat{\lambda})$ if

1. *$\phi_r^{\omega*} \notin \hat{\Phi}^{\omega*}$ or*
2. *$(\hat{\sigma}_1^{\omega},\hat{\sigma}_0^{\omega}) \neq (\sigma_1^{\omega},\sigma_0^{\omega})\ \forall \hat{\sigma}^{\omega} \in \Sigma^{\omega}(\phi_r^*,\hat{\lambda}), \sigma^{\omega} \in \Sigma^{\omega}(\phi_r^*,\lambda^d)$.*

Proof: see appendix B.3.

Lemma 5.6 states that the biological mutants obtain strictly less biological fitness than their peers if (1) the social norm $\phi_r^{\omega*}$ is not an element of the post-mutation cultural equilibrium $\hat{\Phi}^{\omega*}$ or (2) norm population behavior of any element in the behavioral equilibrium at social norm $\phi_r^{\omega*}$ differs for preference distributions λ and $\hat{\lambda}$.

Since the social norm $\phi_r^{\omega*}$ is a cultural equilibrium of proposition 4.3 at preference distribution λ^d, all norm non-holders of the dominant preference type (would) strictly prefer to defect and all norm holders to cooperate, $\theta_s^d\tilde{v}(\phi_r^{\omega*}) < -\Delta m^\omega(\phi_r^{\omega*}) < \theta_s^d\tilde{v}(\phi_r^{\omega*}) + \theta_s^d\tilde{h} + \theta_p^d$. Consequently, cooperating maximizes biological fitness if and only if an individual holds the cooperation norm, $b(n, n, \phi_r^{\omega*}, \phi_r^{\omega*}) > b(1-n, n, \phi_r^{\omega*}, \phi_r^{\omega*})\ \forall n \in \{0, 1\}$.

Suppose biological mutation to $\hat{\lambda}$ alters the social norm or norm population behavior at social norm $\phi_r^{\omega*}$ so that condition 1 or condition 2 of lemma 5.6 holds. In both cases, there exists a cultural equilibrium $\hat{\Phi}^{\omega*}$ after biological mutation that is very close to $\phi_r^{\omega*}$ (see lemma 5.3). Since the cultural equilibrium $\hat{\Phi}^{\omega*}$ is close to $\phi_r^{\omega*}$, cultural evolution reaches it after the biological mutation occurs. For the reminder, consider any possible post-mutation social norm $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$. Since the social norm $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ is close to $\phi_r^{\omega*}$ and preference distribution $\hat{\lambda}$ is close to λ, the cooperation share $\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$ and, consequently, the costs of contribution $-\Delta m^\omega(\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}))$ remain close to $\psi^{\omega*}(\phi_r^{\omega*}, \lambda) = \phi_r^{\omega*}$ and $-\Delta m^\omega(\psi^{\omega*}(\phi_r^{\omega*}, \lambda))$ respectively. These small changes imply that cooperation at preference distribution $\hat{\lambda}$ still maximizes biological fitness if and only if an individual holds the cooperation norm, $b^\omega(n, n, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) > b^\omega(1-n, n, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega)\ \forall n \in \{0, 1\}$. Moreover, all individuals whose approval preferences did not change due to biological mutation (all non-mutants) behave accordingly. Since they previously strictly preferred to cooperate or defect, the small changes in the social norm and contribution costs do not alter their equilibrium behavior. Consequently, they still maximize their biological fitness at preference distribution $\hat{\lambda}$. However, some of the biological mutants must behave differently. If this were not the case, everyone would behave as if only θ^d existed. Given this equilibrium behavior, cultural evolution would instate the social norm $\phi_r^{\omega*}$.[17] The mutants who deviate from biological fitness-maximizing obtain less biological fitness than their peers in situation ω. It follows that

[17]Suppose norm population behavior would mimic that of the λ^d society at $\hat{\phi}^\omega \neq \phi^{\omega*}$. Since $\hat{\phi}^\omega$ is close to $\phi_r^{\omega*}$ and $\phi_r^{\omega*}$ is a cultural equilibrium at λ^d, norm population behavior $(1, 0)$ would imply that society evolves towards $\phi_r^{\omega*}$. Hence, $\hat{\Phi}^{\omega*}$ could not be a cultural equilibrium.

average biological fitness of preference distribution λ outweighs that of $\hat{\lambda}$ at the post-mutation norm and equilibrium behavior in situation ω, $B^{\omega}_{\lambda}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) > B^{\omega}_{\hat{\lambda}}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) \, \forall \hat{\sigma}^{\omega} \in \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})$.

Note that conditions 1 and 2 of lemma 5.6 do not correspond to the negation of conditions 1 and 2 of lemma 5.4. Hence, our results on situations with imperfect social norms $\phi^{\omega *}_r \in (0,1)$ yield no insights into what happens if, after biological mutation, (1) $\hat{\Phi}^{\omega *} \neq \{\phi^{\omega *}\}$ but $\phi^{\omega *} \in \hat{\Phi}^{\omega *}$ or (2) $(\hat{\sigma}^{\omega}_1, \hat{\sigma}^{\omega}_0) \neq (\sigma^{\omega}_1, \sigma^{\omega}_0)$ for some but not all $\hat{\sigma}^{\omega} \in \Sigma^{\omega}(\phi^*_r, \hat{\lambda})$ and $\sigma^{\omega} \in \Sigma^{\omega}(\phi^*_r, \lambda^d)$. The following lemma covers this case.

Lemma 5.7. *Consider any ϕ^*_r of definition 5.6 and $\Lambda_p(\phi^*_r)$ of definition 5.5. For each preference distribution $\hat{\lambda} \in [0,1]^{|\Theta|}$, let $\hat{\Phi}^{\omega *}$ be the cultural equilibrium at $\hat{\lambda}$ s.t. ϕ^*_r is in it's basin of attraction.*

For all $\lambda \in \Lambda_p(\phi^*_r)$, there is some neighborhood U of λ s.t. for all $\hat{\lambda} \in U$ and $\omega \in \{x \in \Omega : \phi^{x*}_r \in (0,1)\}$: If

1. *$\hat{\Phi}^{\omega *} \neq \{\phi^{\omega *}_r\}$ but $\phi^{\omega *}_r \in \hat{\Phi}^{\omega *}$ or*
2. *$(\hat{\sigma}^{\omega}_1, \hat{\sigma}^{\omega}_0) \neq (\sigma^{\omega}_1, \sigma^{\omega}_0)$ for some but not all $\hat{\sigma}^{\omega} \in \Sigma^{\omega}(\phi^*_r, \hat{\lambda})$ and $\sigma^{\omega} \in \Sigma^{\omega}(\phi^*_r, \lambda^d)$,*

then

(i) $B^{\omega}_{\lambda}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) \geq B^{\omega}_{\hat{\lambda}}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega})$ for all $\hat{\phi}^{\omega} \in \hat{\Phi}^{\omega *}$ and $\hat{\sigma}^{\omega} \in \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})$ and

(ii) $\exists \bar{\omega} \in \Omega$ s.t. $B^{\bar{\omega}}_{\lambda}(\hat{\sigma}^{\bar{\omega}}, \hat{\phi}^{\bar{\omega}}) > B^{\bar{\omega}}_{\hat{\lambda}}(\hat{\sigma}^{\bar{\omega}}, \hat{\phi}^{\bar{\omega}})$ for all $\hat{\phi}^{\bar{\omega}} \in \hat{\Phi}^{\bar{\omega} *}$ and $\hat{\sigma}^{\bar{\omega}} \in \Sigma^{\bar{\omega} *}(\hat{\phi}^{\bar{\omega}}, \hat{\lambda})$.

Proof: The lemma is a corollary of lemma B.16 and lemma B.17 in appendix B.3.

The lemma states that in the described case, (i) the biological mutants obtain weakly less biological fitness than their peers in the respective situation, but (ii) there is another situation $\bar{\omega}$ in which they must be obtaining strictly less biological fitness.

By similar reasoning as for lemma 5.6, we can show that, after the biological mutation, cooperation maximizes biological fitness if and only if an individual holds the cooperation norm and all non-mutants behave accordingly. Hence, the non-mutants maximize biological fitness, implying that $B^{\omega}_{\lambda}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) \geq B^{\omega}_{\hat{\lambda}}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) \, \forall \hat{\phi}^{\omega} \in \hat{\Phi}^{\omega *}$ and $\hat{\sigma}^{\omega} \in \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})$.

Moreover, note that condition 2 of lemma 5.7 can only be true if some norm holders and norm non-holders are indifferent between both behavioral routines at social norm $\phi_r^{\omega*}$ and preference distribution $\hat{\lambda}$ (follows from lemma B.7). Similarly, $\phi_r^{\omega*} \in \hat{\Phi}^{\omega*}$ implies that $\phi_r^{\omega*}$ is a rest point of norm dynamics at preference distribution $\hat{\lambda}$. Since, $\phi_r^{\omega*}$ is a cultural equilibrium of proposition 4.3 at λ^d, this can only be true if $(\hat{\sigma}_1^{\omega}, \hat{\sigma}_0^{\omega}) = (1, 0)$ for some $\hat{\sigma}^{\omega} \in \Sigma^{\omega}(\phi_r^{\omega*}, \hat{\lambda})$. Otherwise, norm evolution would not be at rest for any behavioral distribution in the behavioral equilibrium. Moreover, there must also be some individuals who are indifferent between both behavioral routines. Otherwise, norm population behavior in some neighborhood of $\phi_r^{\omega*}$ would correspond to cooperation by all and only norm holders. However, $\phi_r^{\omega*}$ corresponds to a cultural equilibrium for such norm population behavior. Condition 1 and condition 2 both imply that there must be some individuals who are indifferent between both behavioral routines.

Suppose some of the indifferent individuals correspond to norm holders, $\theta_s\tilde{v}(\phi_r^{\omega*}) + \theta_s\tilde{h} + \theta_p = -\Delta m^{\omega}(\phi_r^{\omega*})$. However, this can only be true if the indifferent individuals' norm-based utility benefits are not above the lower bound introduced by the most costly situation with a perfect social norm inducing full cooperation at $\Lambda_p(\phi_r^{\omega*})$ (recall lemma 5.2). This can only occur if the individuals prefer not to cooperate in the most costly situation with a perfect social norm inducing full cooperation at $\Lambda_p(\phi_r^{\omega*})$, $\theta_s\tilde{v}(1) + \theta_s\tilde{h} + \theta_p < \max_{\bar{\omega}\in\Omega}\{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*}) : \phi_r^{\bar{\omega}*} = \psi^{\bar{\omega}*}(1, \hat{\lambda}) = 1\}$. Lemma 5.5 then implies that the mutants obtain less biological fitness than their peers in that situation.

Suppose there are no indifferent norm holders but only norm non-holders, $\theta_s\tilde{v}(\phi_r^{\omega*}) = -\Delta m^{\omega}(\phi_r^{\omega*})$. Note that the situation ω cannot correspond to the one inducing the upper bound on preferences for social approval, $\omega \neq \text{argmin}_{\bar{\omega}\in\Omega}\{\frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*})}{\tilde{v}(\phi_r^{\bar{\omega}*})} : \phi_r^{\bar{\omega}*} \in (0, 1)\}$. Otherwise, condition 2 of definition 5.6 and proposition 4.9 would imply that $\phi_r^{\omega*}$ would not be an element of the post-mutation cultural equilibrium. Hence, the preferences for social approval of the indifferent norm non-holders satisfies $\theta_s = \frac{-\Delta m^{\omega}(\phi_r^{\omega*})}{\tilde{v}(\phi_r^{\omega*})} > \frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*})}{\tilde{v}(\phi_r^{\bar{\omega}*})}$. This can be rearranged to show that the norm non-holders indifferent in situation ω strictly prefer to cooperate in the situation inducing the lowest upper bound on preferences for social approval. Hence, they earn less biological fitness in that situation than their peers. Thus, if condition 1 or condition 2 of lemma 5.7 holds, then there must be some other situation $\bar{\omega}$

for which the biological mutants obtain strictly less biological fitness than their peers.

We can now combine the results of lemma 5.4, lemma 5.5, lemma 5.6, and lemma 5.7 to show that the biological mutants always obtain strictly less biological fitness than their peers after the biological mutation to $\hat{\lambda}$, $B_{\lambda}(\hat{\sigma}, \hat{\phi}) > B_{\hat{\lambda}}(\hat{\sigma}, \hat{\phi}) \, \forall \hat{\phi} \in \prod_{\omega \in \Omega} \hat{\Phi}^{\omega *}$ and $\hat{\sigma} \in \prod_{\omega \in \Omega} \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})$. Lemma 5.4 implies that the biological mutants obtain the same biological fitness as their peers in any situation for which the social norm and equilibrium norm population behavior are not affected. Since $\hat{\lambda} \notin \Lambda_p(\phi_r^*)$, there must be a situation $\omega \in \Omega$ for which either the social norm is altered or equilibrium norm population at social norm $\phi_r^{\omega *}$. If the situation features a perfect social norm $\phi_{\omega}^{\omega *} = 1$ at $\Lambda_p(\phi_r^{\omega *})$, then lemma 5.5 implies that the mutants obtain strictly less biological fitness than their peers in that situation. Alternatively, suppose the situation features an imperfect social norm $\phi_{\omega}^{\omega *} = 1$ at $\Lambda_p(\phi_r^{\omega *})$. In that case, either lemma 5.6 or lemma 5.7 applies to that situation. If the former is the case, the mutants obtain strictly less biological fitness than their peers in that situation. If the latter applies, the mutants obtain weakly less biological fitness than their peers in that situation and strictly less biological fitness than their peers in another situation. Hence, the mutants obtain weakly less biological fitness in all situations and strictly less biological fitness in some situations. It follows that overall, they obtain less biological fitness than their peers, $B_{\lambda}(\hat{\sigma}, \hat{\phi}) > B_{\hat{\lambda}}(\hat{\sigma}, \hat{\phi}) \, \forall \hat{\phi} \in \prod_{\omega \in \Omega} \hat{\Phi}^{\omega *}$ and $\hat{\sigma} \in \prod_{\omega \in \Omega} \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})$. The mutants erode, and biological evolution drives preferences back towards $\Lambda_p(\phi_r^*)$.

Throughout biological evolution, the social norms $\hat{\phi}$ always remain close to ϕ_r^*. Therefore, cultural evolution returns to the social norms ϕ_r^* once preferences return to $\Lambda_p(\phi_r^{\omega *})$. Hence, proposition 5.1 is true. $\Lambda_p(\phi_r^{\omega *})$ is a biological equilibrium for which the social norms ϕ_r^* prevail.

5.3 Discussion

This section has proven the existence of a biological equilibrium for which varying social norms and cooperation levels across situations persist, if the set of situations is sufficiently large and divers. Although the population is potentially heterogeneous regarding preferences, it behaves on aggregate as if it was homogeneous. The respective preference type that quasi-

persists maximizes biological fitness. Preferences for social approval account for the impact of social approval on biological fitness. The weight of social approval on biological fitness ρ and gossip δ increase preferences for social approval and, thus, feasible cooperation. Preferences for self-approval compensate for the inability to overlook the whole extent of actions in terms of social disapproval from hypocrisy.

Suppose individuals would be able to oversee the whole extent, $h = (1 + \delta)\tilde{h}$, then no personal concerns are necessary for the dominant preference type θ^d to maximize biological fitness. Individuals who hold the cooperation norm would then follow their personal norms only to avoid social disapproval from hypocrisy. The complete absence of social disapproval for hypocrisy, $h = \tilde{h} = 0$, implies that individuals who hold the cooperation norm have neither personal nor social incentives to follow their personal norm, $\theta_s^d \tilde{h} + \theta_p^d = 0$. Behavior across both norm populations is equal, implying no differences in average material payoff and social disapproval from social norm violation. Norm evolution is solely driven by conformity and society only reaches cultural equilibria at extreme levels, $\phi^{\omega *} \in \{0, 1\}\ \forall \omega \in \Omega$. The diversity of culture vanishes.

If social disapproval for non-conformity were also absent, then norm evolution would be subject to a random walk. Norm-driven cooperation could only persist if the formation of norms was exogenous or subject to some other mechanism such as institutional pressure (see e.g., Gintis 2003a; Mengel 2008) or conformity bias in social learning (see e.g., Chudek and Joseph Henrich 2011; Joe Henrich and Robert Boyd 1998; Joseph Henrich and Robert Boyd 2001; Michaeli and Spiro 2015; Nordblom and Žamac 2012). The above discussion already highlights the importance of non-conformity in the absence of hypocrisy. However, conformity concerns are crucial even in the presence of social disapproval for hypocrisy. They stabilize perfect social norms and create some robustness concerning preference evolution. In particular, social disapproval for non-conformity ensures that the perfect social norm may remain a cultural equilibrium in the neighborhood of the biological equilibrium.[18]

Securing the persistence of perfect social norms when accounting for biological evolution is particularly crucial, since the definition 5.6 of stability-inducing equilibrium culture builds on inter-dependencies between different situations and cultural equilibria. In particular, for

[18] Appendix A.5 provides an illustrative example.

a cultural equilibrium of an imperfect social norm to occur in some situation, definition 5.6 requires the existence of a cultural equilibrium of a perfect social norm inducing full cooperation at sufficiently large costs. Hence, by stabilizing and enabling the persistence of perfect social norms, social disapproval for conformity in fact enables the persistence of any cooperation-inducing social norms.

6 Concluding Remarks

This paper contributes to the theoretical literature exploring the evolutionary roots of norm-driven cooperation. The results suggest that if norm and preference transmission depends on material and social factors, then an interplay of social disapproval mechanisms can explain the persistence of a diverse culture with varying social norms and cooperation levels across situations. Although in equilibrium preferences are potentially heterogeneous, behavior is as if they were homogeneous. Social disapproval for social norm violation provides individuals with incentives to cooperate at the behavioral level, favors norm evolution at the cultural level, and allows for social approval preferences at the biological level. Social disapproval for non-conformity stabilizes perfect social norms at the cultural level. It favors norm evolution if the social norm is relatively strong. Social disapproval for hypocrisy introduces cooperation incentives at the behavioral and biological levels. Thereby, it enables the heterogeneity in cooperative behavior across individuals and situations. However, it negatively impacts the cultural fitness of individuals with cooperation prescribing personal norms that defect, which can hinder the preservation of social norms if cooperation is very costly. The interplay of these social disapproval mechanisms provides a complementary explanation to assortative matching (e.g., Alger and Weibull 2013, 2016; Mengel 2008) and institutional pressure (e.g. Gintis 2003a; Mengel 2008) for the persistence of norm-driven cooperation. The results also contribute to the literature on gene-culture co-evolution by highlighting how the existing culture shapes the genes that may prevail in equilibrium (e.g., Chudek and Joseph Henrich 2011; Gintis 2003b, 2011; P. Richerson and Rob Boyd 2010; Peter J Richerson et al. 2010).

One of the primary motivations for developing this model was the mutual endogenization of norms and preferences that induce norm adherence. The goal was to draw a more complete

picture of the underlying dynamic system. Our results underscore the importance thereof. When considering norm or preference evolution by itself, disapproval for social norm violation is sufficient to explain large-scale cooperation. However, this is not true when endogenizing both. Moreover, the set of social norms that may prevail as cultural equilibria shrinks when accounting for the evolution of approval preferences. Nevertheless, depending on the purpose of the analysis, it might be sensible to look at cultural evolution by itself. Arguably, the set of situations varies over time (e.g., due to technological changes). The creation and dismantling of situations may occur so fast that biological evolution does not interfere with their respective cultural equilibria.

Furthermore, the findings support insights from Traxler and Spichtig (2011) on the interaction in heterogeneous environments to explain the persistence of empirically observed behavioral patterns. The similarity in the results is apparent despite differences in the setup. We introduced heterogeneous environments as different situations and found that essentially independent situations become interdependent. Traxler and Spichtig (2011) introduce heterogeneous environments through different behavioral equilibria that society coordinates into in the same situation.

Moreover, this paper complements the analysis by Traxler and Spichtig (2011) in another way. We endogenize social norm strength through the distribution of personal norms, whereas Traxler and Spichtig (2011) endogenize it through actual behavior. These two approaches are not mutually exclusive. Its inclusion in our analysis would lead for norm-based utility benefits in figure 1 to be increasing in ψ^ω with the point ϕ^ω being the only exception. As a consequence, the set of Nash equilibria is possibly non-connected, and multiple equilibrium cooperation shares $\psi^{\omega*}$ exist. The underlying reason is similar to the case of decreasing contribution costs discussed in A.3. The cultural equilibria presented in 4 still exist. However, the stability conditions must account for social disapproval for behavioral non-conformity.

The model of this paper builds on some notable assumptions that need further investigation. First, we assumed that the reproduction of norms and preferences depend on social approval. However, how exactly this occurs is left as somewhat of a black box. One possible explanation is that social disapproval is associated with lower material payoff. This per-

spective is in line with traditional approaches from *evolutionary game theory* that consider material payoff as the sole determinant of reproductive fitness. This paper more closely aligns with approaches from *cultural evolutionary models*. We assume that the transmission of traits occurs through complex channels and is biased by social status. However, this leaves unanswered how the weights of reproductive fitness are precisely determined, which is likely a process endogenous to society. Future research in that direction needs to complement this paper.

Another aspect that requires further investigation is the role of communication. Throughout the analysis, we assume that individuals engage in gossip, express disapproval, and share their personal norms. Communication with peers is arguably costly and may provide limited benefits. Therefore, it can be regarded as a public goods dilemma in which individuals must cooperate to sustain cooperation across various other situations. Engaging in gossip about others does not create any incentive problems regarding own optimal behavior. Therefore, we can apply the results of this paper to explain why individuals gossip with peers. We cannot apply this argument to the other two communication dimensions since social disapproval for hypocrisy and non-conformity introduce incentives to misrepresent personal norms.[19] In 2, we argue that a positive probability of being detected when lying and severe material or social costs, in that case, may cause truth-telling to be an alternativeless best-reply. Nevertheless, further work incorporating communication as a behavioral dimension is needed to complement this paper.[20]

Furthermore, future research should also investigate the interaction in more complex environments where situations are possibly interdependent and vary regarding more characteristics than only the material costs of contribution.

[19]Note that expressing social disapproval towards others (for their personal norms and behavior) and communicating one's personal norms coincides to some extent. An individual implicitly shares what she considers morally appropriate whenever she openly disapproves of another individual's personal norm or action. Similarly, by communicating personal views on morally appropriate behavior, she expresses whether she consents to another individual's actions and personal norms.

[20]Models that investigate the signaling of preferences in an evolutionary setting are proposed by Gintis et al. (2001) and Müller and Wangenheim (2019) among others.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond (2019). "Preferences for truth-telling". In: *Econometrica* 87.4, pp. 1115–1153.

Akerlof, George A and Rachel E Kranton (2005). "Identity and the Economics of Organizations". In: *Journal of Economic perspectives* 19.1, pp. 9–32.

Alger, Ingela and Jörgen W Weibull (2013). "Homo moralis—preference evolution under incomplete information and assortative matching". In: *Econometrica* 81.6, pp. 2269–2302.

— (2016). "Evolution and Kantian morality". In: *Games and Economic Behavior* 98, pp. 56–67.

Andreoni, James and B Douglas Bernheim (2009). "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects". In: *Econometrica* 77.5, pp. 1607–1636.

Azar, Ofer H (2004). "What sustains social norms and how they evolve?: The case of tipping". In: *Journal of Economic Behavior & Organization* 54.1, pp. 49–64. ISSN: 0167-2681. DOI: https://doi.org/10.1016/j.jebo.2003.06.001. URL: http://www.sciencedirect.com/science/article/pii/S016726810300221X.

Bašić, Zvonimir and Simone Quercia (2022). "The influence of self and social image concerns on lying". In: *Games and Economic Behavior* 133, pp. 162–169. ISSN: 0899-8256. DOI: https://doi.org/10.1016/j.geb.2022.02.006. URL: https://www.sciencedirect.com/science/article/pii/S0899825622000513.

Bénabou, Roland and Jean Tirole (2006). "Incentives and prosocial behavior". In: *American economic review* 96.5, pp. 1652–1678.

— (2011). "Identity, morals, and taboos: Beliefs as assets". In: *The Quarterly Journal of Economics* 126.2, pp. 805–855.

Bereczkei, Tamas and Andras Csanaky (1996). "Mate choice, marital success, and reproduction in a modern society". In: *Ethology and Sociobiology* 17.1, pp. 17–35.

Bernheim, B Douglas (1994). "A theory of conformity". In: *Journal of political Economy* 102.5, pp. 841–877.

Bester, Helmut and Werner Güth (1998). "Is altruism evolutionarily stable?" In: *Journal of Economic Behavior & Organization* 34.2, pp. 193–209.

Bezin, Emeline (2019). "The economics of green consumption, cultural transmission and sustainable technological change". In: *Journal of Economic Theory* 181, pp. 497–546. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2019.03.005. URL: https://www.sciencedirect.com/science/article/pii/S0022053119300298.

Binmore, Ken and Larry Samuelson (1994). "An economist's perspective on the evolution of norms". In: *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, pp. 45–63.

Bisin, Alberto, Giorgio Topa, and Thierry Verdier (2004). "Cooperation as a transmitted cultural trait". In: *Rationality and Society* 16.4, pp. 477–507.

Bisin, Alberto and Thierry Verdier (2001). "The economics of cultural transmission and the dynamics of preferences". In: *Journal of Economic theory* 97.2, pp. 298–319.

Blau, Peter M (1964). "Exchange and power in social life. New Brunswick". In.

Bowles, Samuel and Herbert Gintis (1998). "The moral economy of communities: Structured populations and the evolution of pro-social norms". In: *Evolution and Human Behavior* 19.1, pp. 3–25.

Boyd, Robert and Peter J Richerson (2005). *The origin and evolution of cultures*. Oxford University Press.

— (1990). "Group selection among alternative evolutionarily stable strategies". In: *Journal of Theoretical Biology* 145.3, pp. 331–342. ISSN: 0022-5193. DOI: https://doi.org/10.1016/S0022-5193(05)80113-4. URL: https://www.sciencedirect.com/science/article/pii/S0022519305801134.

Brekke, Kjell Arne, Snorre Kverndokk, and Karine Nyborg (2003). "An economic model of moral motivation". In: *Journal of public economics* 87.9-10, pp. 1967–1983.

Buss, David M and David P Schmitt (1993). "Sexual strategies theory: an evolutionary perspective on human mating." In: *Psychological review* 100.2, p. 204.

Carbonara, Emanuela, Francesco Parisi, and Georg Von Wangenheim (2008). "Lawmakers as norm entrepreneurs". In: *Review of Law & Economics* 4.3, pp. 779–799.

Chudek, Maciej and Joseph Henrich (2011). "Culture–gene coevolution, norm-psychology and the emergence of human prosociality". In: *Trends in cognitive sciences* 15.5, pp. 218–226.

Cinyabuguma, Matthias, Talbot Page, and Louis Putterman (2005). "Cooperation under the threat of expulsion in a public goods experiment". In: *Journal of public Economics* 89.8, pp. 1421–1435.

Cooter, Robert (1998). "Expressive law and economics". In: *The Journal of Legal Studies* 27.S2, pp. 585–607.

Crawford, Sue ES and Elinor Ostrom (1995). "A grammar of institutions". In: *American political science review*, pp. 582–600.

d'Adda, Giovanna et al. (2020). "Social norms with private values: Theory and experiments". In: *Games and Economic Behavior* 124, pp. 288–304.

Elster, Jon (1989). "Social norms and economic theory". In: *Journal of economic perspectives* 3.4, pp. 99–117.

Fehr, Ernst and Urs Fischbacher (2004). "Social norms and human cooperation". In: *Trends in Cognitive Sciences* 8.4, pp. 185–190. ISSN: 1364-6613. DOI: https://doi.org/10.

1016/j.tics.2004.02.007. URL: http://www.sciencedirect.com/science/article/pii/S1364661304000506.

Figueres, Charles, David Masclet, and Marc Willinger (2013). "Weak moral motivation leads to the decline of voluntary contributions". In: *Journal of Public Economic Theory* 15.5, pp. 745–772.

Gächter, Simon and Ernst Fehr (1999). "Collective action as a social exchange". In: *Journal of economic behavior & organization* 39.4, pp. 341–369.

Geary, David C, Jacob Vigil, and Jennifer Byrd-Craven (2004). "Evolution of human mate choice". In: *Journal of sex research* 41.1, pp. 27–42.

Gintis, Herbert (2003a). "Solving the puzzle of prosociality". In: *Rationality and Society* 15.2, pp. 155–187.

— (2003b). "The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms". In: *Journal of theoretical biology* 220.4, pp. 407–418.

— (2011). "Gene–culture coevolution and the nature of human sociality". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1566, pp. 878–888.

Gintis, Herbert, Eric Alden Smith, and Samuel Bowles (2001). "Costly signaling and cooperation". In: *Journal of theoretical biology* 213.1, pp. 103–119.

Güth, Werner and Menahem Yaari (1992). "An evolutionary approach to explain reciprocal behavior in a simple strategic game". In: *U. Witt. Explaining Process and Change–Approaches to Evolutionary Economics. Ann Arbor*, pp. 23–34.

Guttman, Joel M (2003). "Repeated interaction and the evolution of preferences for reciprocity". In: *The economic journal* 113.489, pp. 631–656.

— (2013). "On the evolution of conditional cooperation". In: *European Journal of Political Economy* 30, pp. 15–34.

Henrich, Joe and Robert Boyd (1998). "The evolution of conformist transmission and the emergence of between-group differences". In: *Evolution and human behavior* 19.4, pp. 215–241.

Henrich, Joseph (2004). "Cultural group selection, coevolutionary processes and large-scale cooperation". In: *Journal of Economic Behavior & Organization* 53.1, pp. 3–35.

Henrich, Joseph and Robert Boyd (2001). "Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas". In: *Journal of theoretical biology* 208.1, pp. 79–89.

Henrich, Joseph and Francisco J Gil-White (2001). "The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission". In: *Evolution and human behavior* 22.3, pp. 165–196.

Hofbauer, Josef and William H Sandholm (2007). "Stable games". In: *2007 46th IEEE Conference on Decision and Control*. IEEE, pp. 3416–3421.

Hofbauer, Josef and William H Sandholm (2009). "Stable games and their dynamics". In: *Journal of Economic theory* 144.4, pp. 1665–1693.

Irons, William (1979). "Cultural and biological success". In: *Evolutionary biology and human social behavior: An anthropological perspective* 284, p. 302.

Lindbeck, Assar, Sten Nyberg, and Jörgen W Weibull (1999). "Social norms and economic incentives in the welfare state". In: *The Quarterly Journal of Economics* 114.1, pp. 1–35.

Mengel, Friederike (2008). "Matching structure and the cultural transmission of social norms". In: *Journal of Economic Behavior & Organization* 67.3-4, pp. 608–623.

Michaeli, Moti and Daniel Spiro (2015). "Norm conformity across societies". In: *Journal of public economics* 132, pp. 51–65.

Mitteldorf, Joshua and David Sloan Wilson (2000). "Population viscosity and the evolution of altruism". In: *Journal of theoretical biology* 204.4, pp. 481–496.

Müller, Stephan and Georg von Wangenheim (2019). "Coevolution of cooperation, preferences, and cooperative signals in social dilemmas". In: *Center for European, Governance, and Economic Development Research, Discussion Paper* 221.

Nordblom, Katarina and Jovan Žamac (2012). "Endogenous Norm Formation Over the Life Cycle-The Case of Tax Morale." In: *Economic Analysis & Policy* 42.2.

Nyborg, Karine (2000). "Homo economicus and homo politicus: interpretation and aggregation of environmental values". In: *Journal of Economic Behavior & Organization* 42.3, pp. 305–322.

— (2018). "Social norms and the environment". In: *Annual Review of Resource Economics* 10, pp. 405–423.

Nyborg, Karine and Mari Rege (2003a). "Does public policy crowd out private contributions to public goods". In: *Public Choice* 115.3, pp. 397–418.

— (2003b). "On social norms: the evolution of considerate smoking behavior". In: *Journal of Economic Behavior & Organization* 52.3, pp. 323–340.

Ostrom, Elinor (2000). "Collective action and the evolution of social norms". In: *Journal of economic perspectives* 14.3, pp. 137–158.

Panebianco, Fabrizio (2016). "The role of persuasion in cultural evolution dynamics". In: *International Review of Economics* 63.3, pp. 233–258.

Poulsen, Anders and Odile Poulsen (2006). "Endogenous preferences and social-dilemma institutions". In: *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, pp. 627–660.

Rabin, Matthew (1995). "Moral preferences, moral constraints, and self-serving biases". In.

Rege, Mari (2004). "Social norms and private provision of public goods". In: *Journal of Public Economic Theory* 6.1, pp. 65–77.

Richerson, Peter and Rob Boyd (2010). "The Darwinian theory of human cultural evolution and gene-culture coevolution". In: *Evolution since Darwin: the first* 150.

Richerson, Peter J, Robert Boyd, and Joseph Henrich (2010). "Gene-culture coevolution in the age of genomics". In: *Proceedings of the National Academy of Sciences* 107.supplement_2, pp. 8985–8992.

Sandholm, William H (2010). *Population games and evolutionary dynamics*. MIT press.

Sethi, Rajiv and Eswaran Somanathan (1996). "The evolution of social norms in common property resource use". In: *The American Economic Review*, pp. 766–788.

Shackelford, Todd K, David P Schmitt, and David M Buss (2005). "Universal dimensions of human mate preferences". In: *Personality and individual differences* 39.2, pp. 447–458.

Tabellini, Guido (2008). "The scope of cooperation: Values and incentives". In: *The Quarterly Journal of Economics* 123.3, pp. 905–950.

Thøgersen, John (2006). "Norms for environmentally responsible behaviour: An extended taxonomy". In: *Journal of environmental Psychology* 26.4, pp. 247–261.

Traxler, Christian (2010). "Social norms and conditional cooperative taxpayers". In: *European Journal of Political Economy* 26.1, pp. 89–103.

Traxler, Christian and Mathias Spichtig (2011). "Social norms and the indirect evolution of conditional cooperation". In: *Journal of Economics* 102.3, pp. 237–262.

Turke, Paul W (1989). "Evolution and the demand for children". In: *Population and development review*, pp. 61–90.

Voss, Thomas (2001). *Game-theoretical perspectives on the emergence of social norms*. na.

Weibull, Jörgen W (1997). *Evolutionary game theory*. MIT press.

Wiederman, M. (1993). "Evolved gender differences in mate preferences: Evidence from personal advertisements". In: *Ethology and Sociobiology* 14, pp. 331–351.

Young, H Peyton (1993). "The evolution of conventions". In: *Econometrica: Journal of the Econometric Society*, pp. 57–84.

— (1996). "The economics of convention". In: *Journal of economic perspectives* 10.2, pp. 105–122.

— (2015). "The evolution of social norms". In: *economics* 7.1, pp. 359–387.

A Supplementary Analysis

A.1 Behavioral Equilibria

Throughout this section, we look at a society that is homogeneous regarding approval preferences with preference type θ. Let $\Delta u_n^\omega := u^\omega(1, n, \psi^\omega, \phi^\omega, \theta) - u^\omega(0, n, \psi^\omega, \phi^\omega, \theta) = \theta_p n + \theta_s \tilde{h} n + \theta_s \tilde{v}(\phi^\omega) - d(\psi^\omega)$ for all $n \in \{0, 1\}$. It becomes apparent that social disapproval for non-conformity does not influence differences in utility. Moreover, we can derive the following lemma.

Lemma A.1. *For any $\psi^\omega \in [0, 1]$ and $\phi^\omega \in [0, 1]$;*

1. $\Delta u_1^\omega \leq 0 \Rightarrow \Delta u_0^\omega < 0$,
2. $\Delta u_0^\omega \geq 0 \Rightarrow \Delta u_1^\omega > 0$.

Suppose an individual who holds the cooperation norm prefers not to contribute in situation ω. In that case, all agents who do not hold the cooperation norm must also prefer to defect. An individual i with $n_i^\omega = 1$ always has additional incentives to cooperate, namely her self-approval and social disapproval from hypocrisy. Analogously, suppose an individual who does not hold the cooperation norm as her personal norm prefers to contribute in situation ω. In that case, any individual who holds the cooperation norm must prefer to do so too. Starting from lemma A.1, we can derive the set of potential Nash equilibria and the conditions under which they exist. The results are presented in table 1. To obtain the equilibrium strategies for Nash equilibria of the second and fourth type, we simply solve the indifference condition $\Delta u_n^\omega = 0$ for n s.t. $\sigma_n^\omega \in (0, 1)$.

A.2 Additional Cultural Equilibria

This section discusses two further cultural equilibria of imperfect social norms that may exist at the λ^d-society. Below, we present these cultural equilibria as well as some graphical illustrations. In the end, we briefly discuss why they cannot prevail when accounting for preference evolution.

	Equilibrium Behavior		Condition for Existence
1.	$\sigma_1^{\omega*}=0,\ \sigma_0^{\omega*}=0,\ \psi^{\omega*}=0$	if	$\theta_s\tilde{v}(\phi^\omega)+\theta_s\tilde{h}+\theta_p\leq-\Delta m^\omega(0)$
2.	$\sigma_1^{\omega*}=\frac{[-\Delta m^\omega]^{-1}(\theta_s\tilde{v}(\phi^\omega)+\theta_s\tilde{h}+\theta_p)}{\phi^\omega}\in(0,1),\ \sigma_0^{\omega*}=0,\ \psi^{\omega*}=\sigma_1^{\omega*}\phi^\omega$	if	$-\Delta m^\omega(0)<\theta_s\tilde{v}(\phi^\omega)+\theta_s\tilde{h}+\theta_p<-\Delta m^\omega(\phi^\omega)$
3.	$\sigma_1^{\omega*}=1,\sigma_0^{\omega*}=0,\psi^{\omega*}=\phi^\omega$	if	$\theta_s\tilde{v}(\phi^\omega)\leq d(\phi^\omega)\leq\theta_s\tilde{v}(\phi^\omega)+\theta_s\tilde{h}+\theta_p$
4.	$\sigma_1^{\omega*}=1,\ \sigma_0^{\omega*}=\frac{[-\Delta m^\omega]^{-1}(\theta_s\tilde{v}(\phi^\omega))}{1-\phi^\omega}\in(0,1),\ \psi^{\omega*}=\phi^\omega+(1-\phi^\omega)\sigma_0^{\omega*}$	if	$-\Delta m^\omega(\phi^\omega)<\theta_s\tilde{v}(\phi^\omega)<-\Delta m^\omega(1)$
5.	$\sigma_1^{\omega*}=1,\ \sigma_0^{\omega*}=1,\ \psi^{\omega*}=1$	if	$-\Delta m^\omega(1)\leq\theta_s\tilde{v}(\phi^\omega)$

Table 1: nash equilibria in ω.

We start with discussing an imperfect social norm equilibrium of partial hypocrisy. Equilibrium behavior in some neighborhood of such a cultural equilibrium $\phi^{\omega*}\in(0,1)$ is

- $\sigma_1^{\omega*}(\phi^\omega)=\frac{[-\Delta m^\omega]^{-1}(\theta_s^d\tilde{v}(\phi^\omega)+\theta_s^d\tilde{h}+\theta_p^d)}{\phi^\omega}\in(0,1)$ and
- $\sigma_0^{\omega*}(\phi^\omega)=0$.

Thus, norm non-holders defect, and some norm holders cooperate. From $\sigma_1^{\omega*}(\phi^\omega)\in(0,1)$ follows that $\theta_s^d\tilde{v}(\phi^\omega)+\theta_s^d\tilde{h}+\theta_p^d=-\Delta m^\omega(\psi^{\omega*})$. For any $\phi^\omega\in(0,1)$ and $\sigma_0^\omega=0$, norm evolution is at rest if and only if $\sigma_1^\omega(\gamma v(\phi^\omega)+\Delta m^\omega(\psi^\omega))-\gamma(1-\sigma_1^\omega)h+\gamma\Delta k(\phi^\omega)=0$. Solving this term yields:

$$\sigma_1^r(\phi^{\omega*})=\frac{\gamma h-\gamma\Delta k(\phi^\omega)}{\gamma h+\gamma v(\phi^\omega)-d(\psi^{\omega*})}.$$

The above indicates the share of cooperators among norm holders for which norm evolution is at rest. Equilibrium behavior at $\phi^{\omega*}$ yields norm evolution to be at rest if $\sigma_1^{\omega*}(\phi^{\omega*})=\sigma_1^r(\phi^{\omega*})$. For illustration purposes, we look at the case of $h<\Delta k(\phi^{\omega*})$. We can easily show that $\sigma^{r'}>0$. In addition, $h<\Delta k(\phi^{\omega*})$ only if $\Delta k(\phi^{\omega*})>0$ and, thus, $\phi^{\omega*}>\frac{1}{2}$. Moreover, $\sigma_1^r(\phi^\omega)\in(0,1)$ if and only if $\gamma v(\phi^\omega)-\theta_p^d-\theta_s^d\tilde{h}-\theta_s^d\tilde{v}(\phi^\omega)<-\gamma\Delta k(\phi^\omega)<0$. From $\gamma v(\phi^\omega)<\theta_p^d+\theta_s^d\tilde{h}+\theta_s^d\tilde{v}(\phi^\omega)=-\Delta m^\omega(\psi^{\omega*})$ follows that

- $\sigma^{\omega*}(\phi^\omega)<\sigma_1^r(\phi^\omega)\Rightarrow\dot{\phi}^\omega>0$ and

- $\sigma^{\omega*}(\phi^\omega) > \sigma_1^r(\phi^\omega) \Rightarrow \dot{\phi}^\omega < 0$.

Consider the graphical illustration in figure 4 a. At social norm $\phi^{\omega*}$, the cooperation share among norm holders is such that norm evolution is at rest, $\sigma_1^{\omega*}(\phi^\omega) = \sigma_1^r(\phi^\omega)$. A small increase in the social norm to $\phi^\omega > \phi^{\omega*}$ yields $\sigma_1^{\omega*}(\phi^\omega)$ above $\sigma_1^r(\phi^\omega)$. Consequently, the social norm weakens at ϕ^ω, and society returns to $\phi^{\omega*}$. $\phi^{\omega*}$ is a stable rest point. By similar reasoning, the intersection $\phi^{\omega u}$ indicates an unstable rest point.

For the existence of an imperfect social norm equilibrium of partial hypocrisy, we need that social disapproval for either hypocrisy or non-conformity exists. Otherwise, norm evolution is only at rest for some $\sigma_1^{\omega*} \in (0,1)$ if $\theta_s^d \tilde{v}(\phi^\omega) + \theta_p^d = -\Delta m^\omega(\sigma_1^\omega \phi^\omega) = \gamma v(\phi^\omega) = \gamma(1+\delta)\tilde{v}(\phi^\omega)$. $\theta_p^d \geq 0$ implies that $\theta_s^d \leq \gamma(1+\delta)$. A small increase in the social norm to $\hat{\phi}^\omega > \phi^\omega$ yields $\theta_s^d \tilde{v}(\hat{\phi}^\omega) + \theta_p^d = -\Delta m^\omega(\sigma_1^\omega \tilde{v}(\hat{\phi}^\omega)) \leq \gamma(1+\delta)\tilde{v}(\hat{\phi}^\omega)$. Social disapproval for social norm violation outweighs material costs, and the social norm is either at rest or increases further.

Next, we investigate imperfect social norm equilibria of social pressure. Equilibrium behavior in some neighborhood of such a cultural equilibrium $\phi^{\omega*} \in (0,1)$ is

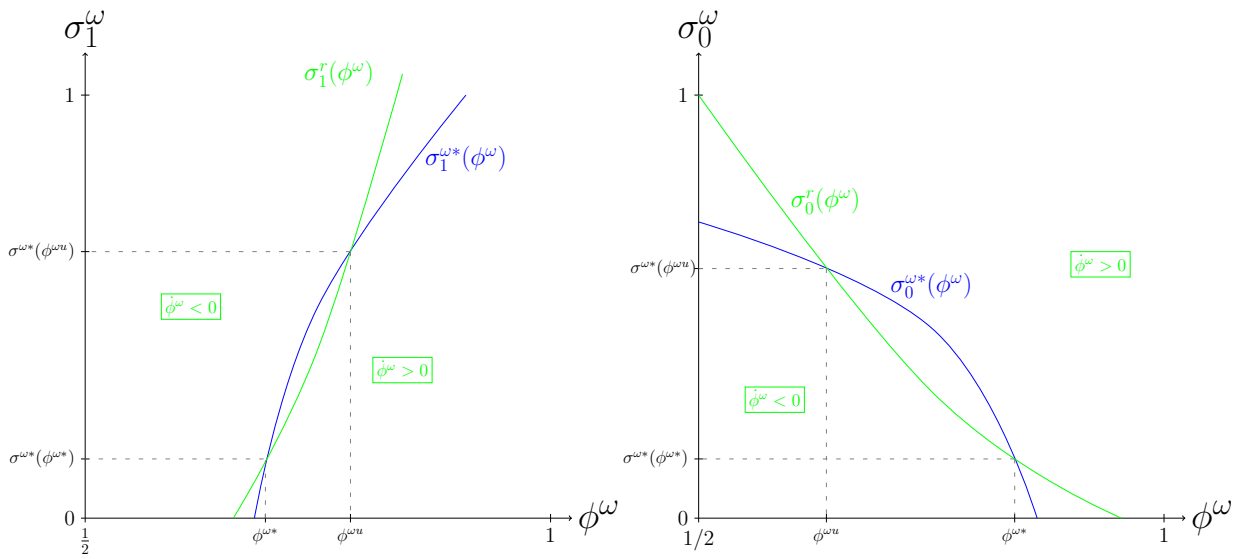
- $\sigma_1^{\omega*}(\phi^\omega) = 1$ and
- $\sigma_0^{\omega*}(\phi^\omega) = \frac{[-\Delta m^\omega]^{-1}(\theta_s^d \tilde{v}(\phi^\omega)) - \phi^\omega}{1-\phi^\omega}$.

Thus, all norm holders and some norm non-holders cooperate. From $\sigma_0^{\omega*}(\phi^\omega) \in (0,1)$ follows that $\theta_s^d \tilde{v}(\phi^\omega) = -\Delta m^\omega(\psi^{\omega*})$. Consider any social norm $\phi^\omega \in (0,1)$. Given $\sigma_1^\omega = 1$, the cooperation share among norm non-holders for which norm evolution is at rest is:

$$\sigma_0^r(\phi^\omega) := 1 + \frac{\gamma \Delta k(\phi^\omega)}{\gamma v(\phi^\omega) - \theta_s^d \tilde{v}(\phi^\omega)}.$$

The social norm $\phi^{\omega*}$ is a cultural equilibrium only if $\sigma_0^r(\phi^{\omega*}) = \sigma_0^{\omega*}(\phi^{\omega*})$. For illustration purposes, we focus on $\phi^{\omega*} > \frac{1}{2}$ in the following. We can easily derive that $\sigma_0^{r'} < 0$ for all $\phi^\omega > \frac{1}{2}$. In addition, $\phi^{\omega*} > \frac{1}{2}$ implies that $\Delta k(\phi^{\omega*}) > 0$. Hence, $\sigma_0^r(\phi^\omega) \in (0,1)$ only if $\gamma v(\phi^\omega) < \theta_s^d \tilde{v}(\phi^\omega)$. Moreover, from $\gamma v(\phi^\omega) < \theta_s^d \tilde{v}(\phi^\omega) = -\Delta m^\omega(\psi^{\omega*})$ follows that

- $\sigma^{\omega*}(\phi^\omega) > \sigma_0^r(\phi^\omega) \Rightarrow \dot{\phi}^\omega > 0$ and



(*a*) partial hypocrisy (*b*) social pressure

Figure 4: additional cultural equilibria.

- $\sigma^{\omega*}(\phi^\omega) < \sigma_0^r(\phi^\omega) \Rightarrow \dot{\phi}^\omega < 0.$

Intersection $\phi^{\omega*}$ in figure 4 (b) constitutes a cultural equilibrium, whereas intersection $\phi^{\omega u}$ is an unstable rest point.

For the existence of an imperfect social norm equilibrium that induces partial cooperation among norm non-holders, social disapproval for non-conformity must exist. Otherwise, a rest point $\phi^\omega \in (0,1)$ s.t. $\sigma_0^{\omega*} \in (0,1)$ must satisfy $\gamma v(\phi^\omega) = -\Delta m^\omega(\psi^{\omega*}) = \theta_s^d \tilde{v}(\phi^\omega)$. However, for any other $\hat{\phi}^\omega \in (0,1)$ close to ϕ^ω it holds that $\theta_s^d \tilde{v}(\hat{\phi}^\omega) = -\Delta m^\omega(\hat{\psi}^{\omega*}) = \gamma v(\phi^\omega)$. Thus, a small increase in the social norm from ϕ^ω to $\hat{\phi}^\omega$ leaves norm evolution at rest.

Lastly, we argue why the discussed cultural equilibria cannot prevail when accounting for biological evolution. Although we focus on a specific example, the underlying reasoning applies in similar fashion to any other cultural equilibrium $\phi^{\omega*} \in (0,1)$ that induces $\sigma_n^{\omega*} \in (0,1)$ for some $n \in \{0,1\}$. Consider a cultural equilibrium $\phi^{\omega*} \in (\frac{1}{2},1)$ such that $\sigma_0^{\omega*} \in (0,1)$ and $\sigma_1^{\omega*} = 1$ at λ^d. Assume that there is some λ^d and social norm $\phi^{\omega*}$. Any preference distribution λ that mimics λ^d in terms of behavior must consist of sufficiently many norm non-holders who cooperate and who do not cooperate at $\phi^{\omega*}$:

- $\sum_{\theta \in \text{supp}(\lambda) \wedge \theta_s \geq \theta_s^d} \lambda_\theta \geq \sigma_0^{\omega*}$ and

- $\sum_{\theta \in \text{supp}(\lambda) \wedge \theta_s \leq \theta_s^d} \lambda_\theta \geq 1 - \sigma_0^{\omega *}$.

Moreover, the above discussion indicates that $\phi^{\omega *}$ is a cultural equilibrium only if $\sigma_0^{\omega *'} < \sigma_0^{r'} < 0$. This is true at λ only if there exist some individuals i for whom $\theta_{si} = \theta_s^d$. If there is no such individual, all individuals strictly prefer to do what they do at the current environment. A small change in the social norm does not alter their optimal behavior such that $\sigma_0^{\omega *'} = 0$. Thus, for all $\lambda \in \lambda^d$, $\sum_{\theta \in \{x \in \text{supp}(\lambda) : x_s = \theta_s^d\}} \lambda_\theta > 0$. Thus, λ^d is not a closed set and, therefore, does not satisfy the formal definition of an asymptotically stable set (see Weibull 1997). Nevertheless, suppose that society experiences random walk across all elements in λ^d. Throughout, the share of individuals for which $\theta_s = \theta_s^d$ varies. Eventually, mutation yields some $\hat{\lambda} \notin \lambda^d$ at which no such preference type is present. The social norm $\phi^{\omega *}$ is not a cultural equilibrium anymore, and society coordinates into some other cultural equilibrium $\hat{\phi}^\omega$. Due to the imitative nature of biological evolution, biological evolution cannot yield a revival of preference types for which $\theta_s = \theta_s^d$ if they are extinct. Hence, society does not return to λ^d, implying that it is not a biological equilibrium.

A.3 Decreasing Costs of Contribution

We assumed that $-\Delta m^\omega$ is an increasing function throughout the analysis. In the following, we discuss how the results change if we relax this assumption. We focus on the implications for behavioral and cultural equilibria. In particular, we show that in any situation ω for which $-\Delta m^\omega$ is decreasing, there can only exist cultural equilibria of an absent or perfect social norm. These cultural equilibria prevail when accounting for biological evolution by similar reasoning as under increasing costs of contribution.

Throughout, we consider a society with preference distribution λ^d. First, we discuss how the behavioral results change. Consider the graphical representation in figure 5. Whenever $NU^\omega(\psi^\omega) > -\Delta m^\omega(\psi^\omega)$, there is some individual who prefers to cooperate but currently defects. Hence, ψ^ω increases. Analogously, $NU^\omega(\psi^\omega) < -\Delta m^\omega(\psi^\omega)$ implies that ψ^ω decreases. Thus, there are three stable rest points in figure 5: (1) $\psi^\omega = 0$, (2) $\psi^\omega = \phi^\omega$, and (3) $\psi^\omega = 1$. In addition, there are two Nash equilibria that constitute unstable rest points of the dynamic system: the intersections of $NU^\omega(\psi^\omega)$ and $-\Delta m^\omega(\psi^\omega)$ in the horizontal segments

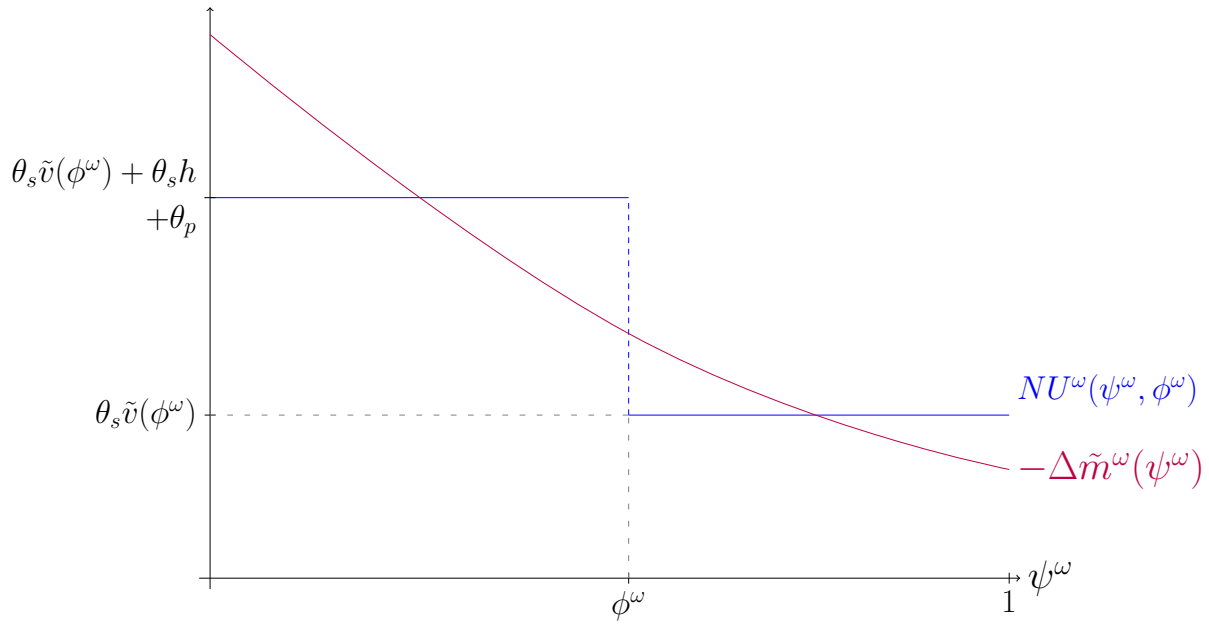


Figure 5: equilibrium behavior under decreasing costs of contribution.

of $NU^\omega(\psi^\omega)$. Hence, if $-\Delta m^\omega$ is a decreasing function, there are possibly multiple Nash equilibria. In addition, any Nash equilibrium is asymptotically stable if and only if it is a strict Nash equilibrium.

It follows that any cultural equilibrium in a situation ω for which $-\Delta m^\omega$ is decreasing must be a cultural equilibrium of (1) an absent social norm, (2) an imperfect social norm s.t. $\psi^{\omega*} = \phi^{\omega*} \in (0, 1)$, or (3) a perfect social norm s.t. $\psi^{\omega*} = \phi^{\omega*} = 1$.

Investigating the existence of such cultural equilibria is similar to the analysis for increasing cost curves. First, the proof of proposition 4.4 can be applied to show that an absent social norm equilibrium always exists. Second, a perfect social norm equilibrium exists only if $\psi^{\omega*} = 1$ at $\phi^\omega = 1$, which for $-\frac{\mathrm{d}\Delta m^\omega(\psi^\omega)}{\mathrm{d}\psi^\omega} < 0$ requires that $\theta_s\tilde{v}(1) + \theta_s\tilde{h} + \theta_p > -\Delta m^\omega(1)$. Under this condition, the proof of proposition 4.5 can be slightly adjusted to show that $\phi^\omega = 1$ is a cultural equilibrium if $-\Delta m^\omega(1) < \max\{\gamma v(1) + \gamma\Delta k(1), \theta_s\tilde{v}(1)\}$. Lastly, we can employ insights from the proof of proposition 4.7 to show that $\phi^\omega \in (0, 1)$ is a cultural equilibrium only if $\gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi^\omega} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi^\omega}) \leq -\frac{\mathrm{d}\Delta m^\omega(\phi^\omega)}{\mathrm{d}x}|_{x=\phi^\omega}$. However, $\frac{\mathrm{d}v(x)}{\mathrm{d}x} > 0$, $\frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x} > 0$, and $-\frac{\mathrm{d}\Delta m^\omega(\phi^\omega)}{\mathrm{d}x} < 0$ imply that this condition can never be satisfied. Figure 6 illustrates this argument graphically. Thus, in any situation where $-\Delta m^\omega$ is decreasing, society reaches a cultural equilibrium of an absent or a perfect social norm.

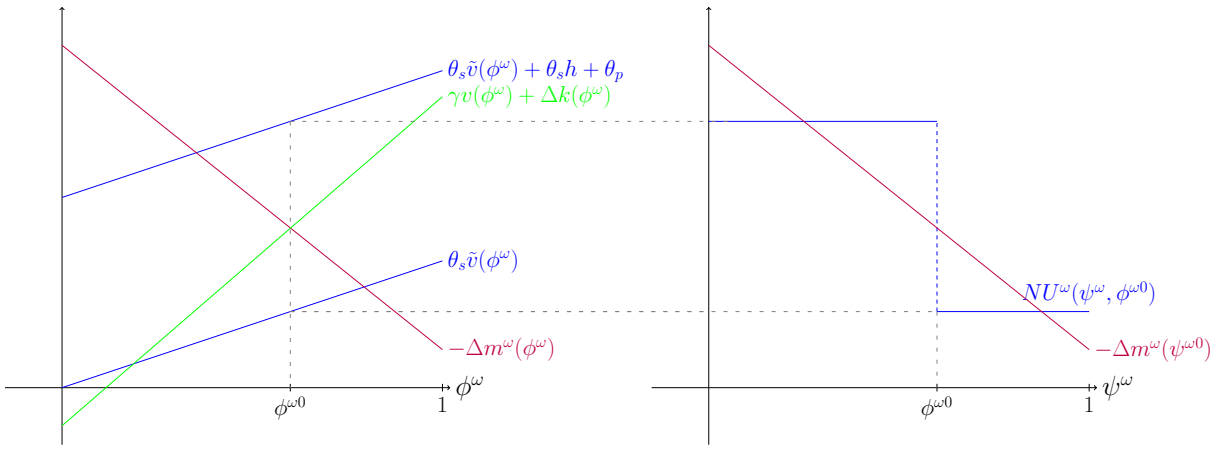


Figure 6: imperfect social norm under decreasing costs of contribution.

The results of section 5 can easily be extended to show the social norms ϕ_r^* satisfying definition 5.6 may also feature perfect social norm and absent social norm equilibria in situations where $-\Delta m^\omega$ is decreasing.

A.4 The Weight of Social Approval on Reproductive Fitnesses

Section 5 introduces the assumption that the weight of social approval on cultural fitness γ is larger than that on biological fitness ρ. This section discusses the consequences of relaxing this assumption. In particular, we investigate different cultural equilibria if approval preferences are distributed according to λ^d and $\rho > \gamma$. Generally speaking, the results of section 4 do not change. However, the support of the different cultural equilibria does. Our following discussion focuses on cultural equilibria of (1) a perfect social norm inducing partial cooperation and (2) an imperfect social norm.

A cultural equilibrium of a perfect social norm inducing partial cooperation can only exist at preference distribution λ^d if $\rho v(1) + \rho h < \gamma v(1) + \gamma \Delta k(1)$. To see this, consider the contrary such that: $\rho v(1) + \rho h \geq \gamma v(1) + \gamma \Delta k(1) \Rightarrow \theta_s^d \tilde{v}(1) + \theta_s^d \tilde{h} + \theta_p^d \geq \gamma v(1) + \gamma \Delta k(1)$. For any cost curve $-\Delta m^\omega$ such that $\psi^{\omega *}(1, \lambda^d) \in (0, 1)$, $-\Delta m^\omega(\psi^{\omega *}) = \theta_s^d \tilde{v}(1) + \theta_s^d \tilde{h} + \theta_p^d > \gamma v(1) + \gamma \Delta k(1) - (1 - \psi^{\omega *})h$. Thus, the perfect social norm cannot be dynamically stable. Hence, $\rho v(1) + \rho h < \gamma v(1) + \gamma \Delta k(1)$ is a necessary condition for the existence of perfect social norm equilibria inducing partial cooperation. Clearly, the set of triplets $(v(1), h, \Delta k(1))$ for which the condition holds is larger if $\gamma > \rho$ rather than $\gamma < \rho$. Hence, we argue that

$\gamma > \rho$ favors the existence of a cultural equilibrium of a perfect social norm inducing partial cooperation.

Next, consider a cultural equilibrium of an imperfect social norm. Such a cultural equilibrium exists only if for some $\phi^\omega \in (0,1)$, $\rho v(\phi^\omega) < \gamma v(\phi^\omega) + \gamma \Delta k(\phi^\omega) < \rho v(\phi^\omega) + \rho h$. We obtain this inequality by substituting for θ_p^d and θ_s^d in the behavioral conditions of proposition 4.3. If $\rho < \gamma$, then there must exist such a ϕ^ω. From $\rho v(0) > \gamma v(0) + \gamma \Delta k(0)$ and $\rho v(\frac{1}{2}) < \gamma v(\frac{1}{2}) + \gamma \Delta k(\frac{1}{2})$ follows that $\rho v(\phi^\omega)$ intersects $\gamma v(\phi^\omega) + \gamma \Delta k(\phi^\omega)$ at some $x \in (0, \frac{1}{2})$ from below. At this intersection x, $\gamma v(x) + \gamma \Delta k(x) < \rho v(x) + \rho h$. Thus, the inequality holds for all social norms ϕ^ω slightly above x. If $\rho \geq \gamma$, then it depends on the specifications of k, v, and h, whether there exists a ϕ^ω that satisfies the inequality. Thus, we must impose some additional structure on the social disapproval mechanisms to ensure that cultural equilibria of imperfect social norms exist.

A.5 Notes on Social Disapproval for Non-Conformity

This section elaborates on the role of social disapproval for non-conformity. In the following, we briefly discuss the consequences of the absence of social disapproval for non-conformity for perfect social norm equilibria inducing complete cooperation at λ^d when endogenizing preference formation. Recall that the existence of a cultural equilibrium of a perfect social norm inducing full cooperation enables the existence of a cultural equilibrium of an imperfect social norm. Hence, by showing the above, we indirectly illustrate the consequences for cooperation-inducing norms as a whole.

In the absence of social disapproval for non-conformity, perfect social norms are no longer stabilized and potentially turn into large sets for small mutations in approval preferences. These sets are potentially unstable. Below we illustrate this point by presenting a specific example. Therefore, suppose that $\Delta k(x) = 0 \; \forall x$. Consider the λ^d society and some cultural equilibrium $\phi^{\omega *} = 1$ s.t. $\min\{\theta_p^d + \theta_s^d \tilde{h} + \theta_s^d \tilde{v}(1), \gamma v(1)\} > -\Delta m^\omega(1)$. $\phi^{\omega *} = 1$ is a cultural equilibrium only if in some close neighborhood of $\phi^{\omega *} = 1$ some non-holders defect. Otherwise, for all social norms in some close neighborhood of $\phi^{\omega *} = 1$, there is no difference in behavior across both norm holder populations, implying both populations ob-

tain the same cultural fitness on average, and norm evolution is at rest. Assume that some preference type θ appears for which $\theta_s > \gamma(1+\delta)$. Hence, all non-holders of this preference type cooperate in the neighborhood of $\phi^{\omega*}$. As long as some individual of preference type θ^d exists, cooperation in the neighborhood of $\phi^{\omega*}$ among norm non-holders is incomplete and the perfect social norm a cultural equilibrium. Formally, if ϕ^ω is close to 1, then $\sigma_0^{\omega*} < 1$, $\sigma_1^{\omega*} = 1$, and $\gamma v(\phi^\omega)\} > -\Delta m^\omega(1) \geq -\Delta m^\omega(\Psi^{\omega*})$, and, thus, $\dot{\phi}^\omega > 0$. In the cultural and behavioral equilibrium, all individuals hold the cooperation norm and cooperate. Hence, preference evolution is at rest. Random walk of approval preferences may eventually lead to the extinction of θ^d by chance. If this happens, all individuals cooperate at $\phi^{\omega*}$ and some neighborhood U. In particular, all individuals cooperate for all $\phi^\omega \geq [\tilde{v}]^{-1}(\frac{-\Delta m^\omega(1)}{\theta_s})$. The set $\Phi^\omega = [[\tilde{v}]^{-1}(\frac{-\Delta m^\omega(1)}{\theta_s}), 1]$ consists of only rest points. Consider some small cultural mutation to $\hat{\phi}^\omega < [\tilde{v}]^{-1}(\frac{-\Delta m^\omega(1)}{\theta_s})$ at the lower bound of Φ^ω. We can easily show that $-\Delta m^\omega(\hat{\phi}^\omega) < \theta_s \tilde{v}(\hat{\phi}^\omega) < -\Delta m^\omega(1)$ implying that $\sigma_0^{\omega*} \in (0,1)$, $\sigma_1^{\omega*} = 1$, and $-\Delta m^\omega(\psi^{\omega*}) = \theta_s \tilde{v}(\hat{\phi}^\omega)$. From $\theta_s \tilde{v}(1) > \gamma v(1)$ follows that $\theta_s \tilde{v}(x) > \gamma v(x)$ for all $x \in [0,1]$. Hence, norm dynamics at $\hat{\phi}^\omega$ are $\dot{\hat{\phi}}^\omega = \hat{\phi}^\omega(1-\hat{\phi}^\omega)[(1-\sigma_0^{\omega*})(\gamma v(\hat{\phi}^\omega)+\Delta m^\omega(\psi^{\omega*}))] = \hat{\phi}^\omega(1-\hat{\phi}^\omega)[(1-\sigma_0^{\omega*})(\gamma v(\hat{\phi}^\omega)-\theta_s \tilde{v}(\hat{\phi}^\omega))] < 0$. The set Φ^ω is unstable, and society evolves to a cultural equilibrium that does not induce complete cooperation.

B Proofs and Additional Formal Results

This section provides the formal proofs of our analysis as well as additional formal results not presented in the main text.

B.1 Behavior

Lemma B.1. *For all $\omega \in \Omega$, $\lambda \in \Theta^{[0,1]}$, and $\phi^\omega \in [0,1]$, the set of Nash equilibria is convex and asymptotically stable.*

Proof. Hofbauer and Sandholm (2007, 2009) show that the set of Nash equilibria is convex and asymptotically stable for any stable game. Consider any two possible strategy distributions $\hat{\sigma}^\omega$ and $\check{\sigma}^\omega$ with the corresponding cooperation shares $\hat{\psi}$ and $\check{\psi}$. The game is stable if $\phi^\omega \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta(\hat{\sigma}_{1,\theta} - \check{\sigma}_{1,\theta})(u^\omega(1,1,\hat{\psi}^\omega,\phi^\omega) - u^\omega(0,1,\hat{\psi}^\omega,\phi^\omega) - u^\omega(1,1,\check{\psi}^\omega,\phi^\omega) + u^\omega(0,0,\check{\psi}^\omega,\phi^\omega)) + (1-\phi) \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta(\hat{\sigma}_{0,\theta} - \check{\sigma}_{0,\theta})(u^\omega(1,n,\hat{\psi}^\omega,\phi^\omega) - u^\omega(0,n,\hat{\psi}^\omega,\phi^\omega) - u^\omega(1,n,\check{\psi}^\omega,\phi^\omega) + u^\omega(0,n,\check{\psi}^\omega,\phi^\omega)) = \phi \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta(\hat{\sigma}_{1,\theta} - \check{\sigma}_{1,\theta})(-\Delta m^\omega(\check{\psi}^\omega) + \Delta m^\omega(\hat{\psi}^\omega)) + (1-\phi) \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta(\hat{\sigma}_{0,\theta} - \check{\sigma}_{0,\theta})(-\Delta m^\omega(\check{\psi}^\omega) + \Delta m^\omega(\hat{\psi}^\omega)) = (\hat{\psi}^\omega - \check{\psi}^\omega)(-\Delta m^\omega(\check{\psi}^\omega) + \Delta m^\omega(\hat{\psi}^\omega)) \leq 0$. Note that this condition is always satisfied, since $-\Delta m^\omega(\cdot)$ is increasing. Thus, the game is stable, which suffices to proof the lemma. □

Lemma B.2. *For all $\omega \in \Omega$, $\lambda \in \Theta^{[0,1]}$, $\phi^\omega \in [0,1]$, $n \in \{0,1\}$, and $\theta \in \text{supp}(\lambda)$, $\sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$ implies:*

1. *$\sigma^\omega_{n,\theta} = 1$ if $n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi^\omega) > -\Delta m^\omega(\phi^\omega \sigma^\omega_1 + (1-\phi^\omega)\sigma^\omega_0)$ and*
2. *$\sigma^\omega_{n,\theta} = 0$ if $n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi^\omega) < -\Delta m^\omega(\phi^\omega \sigma^\omega_1 + (1-\phi^\omega)\sigma^\omega_0)$.*

Proof. Consider any $\sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$. Consider the difference in utilities from cooperation and defection for individuals with personal norm n and preference type θ: $\Delta u^\omega_{n,\theta} = u^\omega(1,n,\psi^\omega,\phi^\omega,\theta) - u^\omega(0,n,\psi^\omega,\phi^\omega,\theta) = n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi^\omega) + \Delta m^\omega(\phi^\omega \sigma^\omega_1 + (1-\phi^\omega)\sigma^\omega_0)$. $n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi^\omega) > -\Delta m^\omega(\phi^\omega \sigma^\omega_1 + (1-\phi^\omega)\sigma^\omega_0)$ implies that cooperation is strictly preferred to defection. Thus, $\sigma^\omega_{n,\theta} = 1$ must be true in any Nash equilibrium. Analogously, $n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi^\omega) < -\Delta m^\omega(\phi^\omega \sigma^\omega_1 + (1-\phi^\omega)\sigma^\omega_0)$ implies that defecting is strictly preferred. Thus, $\sigma^\omega_{n,\theta} = 0$ must be true in any Nash equilibrium. □

Lemma B.3. *For all* $\omega \in \Omega$, $\lambda \in \Theta^{[0,1]}$, $\phi^\omega \in [0,1]$, *and* $\hat{\sigma}^\omega, \check{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$, $\phi^\omega \hat{\sigma}_1^\omega + (1 - \phi^\omega)\hat{\sigma}_0^\omega = \phi^\omega \check{\sigma}_1^\omega + (1 - \phi^\omega)\check{\sigma}_0^\omega$.

Proof. Assume by contradiction that $\exists \hat{\sigma}^\omega, \check{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$ s.t. $\hat{\psi}^\omega := \phi^\omega \hat{\sigma}_1^\omega + (1 - \phi^\omega)\hat{\sigma}_0^\omega > \check{\psi}^\omega := \phi^\omega \check{\sigma}_1^\omega + (1 - \phi^\omega)\check{\sigma}_0^\omega$. It follows that $-\Delta m^\omega(\hat{\psi}^\omega) > -\Delta m^\omega(\check{\psi}^\omega)$. For all $(n, \theta) \in \{0,1\} \times \text{supp}(\lambda)$, $(\theta_s \tilde{v}(\phi^\omega) + n(\theta_s \tilde{h} + \theta_p) \geq -\Delta m^\omega(\hat{\psi}^\omega) \Rightarrow \theta_s \tilde{v}(\phi^\omega) + n(\theta_s \tilde{h} + \theta_p) > -\Delta m^\omega(\check{\psi}^\omega)) \Rightarrow \hat{\sigma}^\omega_{n,\theta} \leq \check{\sigma}^\omega_{n,\theta} \Rightarrow \hat{\psi}^\omega \leq \check{\psi}^\omega$. We have reached a contradiction. □

Lemma B.4. *For all* $\omega \in \Omega$, $\lambda \in \Theta^{[0,1]}$, $\check{\theta}, \hat{\theta} \in \text{supp}(\lambda)$, $\phi^\omega \in [0,1]$, *and* $\check{n}, \hat{n} \in \{0,1\}$, $(\check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\phi^\omega) > \hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\phi^\omega)) \Rightarrow (\sigma^\omega_{\check{n},\check{\theta}} \geq \sigma^\omega_{\hat{n},\hat{\theta}}\ \forall \sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda))$.

Proof. Assume by contradiction that for some $\omega \in \Omega$, $\lambda \in [0,1]^{|\Theta|}$ and $\phi^\omega \in [0,1]$ there exists $\sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$ s.t. $\check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\phi^\omega) > \hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\phi^\omega)$ and $\sigma^\omega_{\check{n},\check{\theta}} < \sigma^\omega_{\hat{n},\hat{\theta}}$ for some $\check{n}, \hat{n} \in \{0,1\}$ and $\check{\theta}, \hat{\theta} \in \text{supp}(\lambda)$. $\sigma^\omega_{\check{n},\check{\theta}} < \sigma^\omega_{\hat{n},\hat{\theta}} \Rightarrow \sigma^\omega_{\hat{n},\hat{\theta}} > 0 \Rightarrow \check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\phi^\omega) \geq -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda)) \Rightarrow \hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\phi^\omega) > -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda)) \Rightarrow \sigma^\omega_{\hat{n},\hat{\theta}} = 1 \Rightarrow \sigma^\omega_{\hat{n},\hat{\theta}} \geq \sigma^\omega_{\check{n},\check{\theta}}$. We have reached a contradiction. □

Lemma B.5. *For all* $\omega \in \Omega$ *and* $\lambda \in \Theta^{[0,1]}$, $\psi^{\omega*}(\phi^\omega, \lambda)$ *is non-decreasing in* ϕ^ω.

Proof. Consider any $\lambda \in \Theta^{[0,1]}$ and $\omega \in \Omega$. We have to show that $x > y \Rightarrow \psi^{\omega*}(x, \lambda) \geq \psi^{\omega*}(y, \lambda)$. Assume by contradiction that $x > y$ and $\psi^{\omega*}(x, \lambda) < \psi^{\omega*}(y, \lambda)$. $\psi^{\omega*}(x, \lambda) < \psi^{\omega*}(y, \lambda) \Rightarrow -\Delta m^\omega(\psi^{\omega*}(x, \lambda)) < -\Delta m^\omega(\psi^{\omega*}(y, \lambda))$. $n(\theta_p + \theta_s \tilde{h}) + \theta \tilde{v}(y) \geq -\Delta m^\omega(\psi^{\omega*}(y, \lambda)) \Rightarrow n(\theta_p + \theta_s \tilde{h}) + \theta \tilde{v}(x) > -\Delta m^\omega(\psi^{\omega*}(x, \lambda))$. It follows that $\hat{\sigma}^\omega_{n,\theta} \geq \tilde{\sigma}^\omega_{n,\theta} \forall \hat{\sigma}^\omega \in \Sigma^{\omega*}(x, \lambda), \tilde{\sigma}^\omega \in \Sigma^{\omega*}(y, \lambda), n \in \{0,1\}, \theta\, \text{supp}(\lambda)$. Hence, $\psi^\omega(x, \lambda) \geq \psi^\omega(y, \lambda)$. We have reached a contradiction. □

Lemma B.6. *For all* $\omega \in \Omega$, $\lambda \in \Theta^{[0,1]}$, $\phi^\omega \in [0,1]$, *and* $\sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$, $\sigma_0^\omega \leq \sigma_1^\omega$.

Proof. Let $y = \sum_{\theta \in \hat{\Theta}} \lambda_\theta$, where $\hat{\Theta} := \{x \in \text{supp}(\lambda) : x_s \tilde{v}(\phi^\omega) \geq -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda))\}$. The share of norm non-holders who strictly prefer to defect is given by $1 - y$. Therefore, $1 - y \leq 1 - \sigma_0^\omega$ and $y > \sigma_0^\omega$. Let $z = \sum_{\theta \in \check{\Theta}} \lambda_\theta$, where $\check{\Theta} := \{x \in \text{supp}(\lambda) : x_s \tilde{v}(\phi^\omega) + x_s \tilde{h} + x_p > -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda))\}$. The share of norm holders who strictly prefer to cooperate is given by z. Thus, $\sigma_1^\omega \geq z$. $(\theta_s \tilde{v}(\phi^\omega) \geq -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda)) \Rightarrow \theta_s \tilde{v}(\phi^\omega) + \theta_s \tilde{h} + \theta_p > -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda))) \Rightarrow (\theta \in \hat{\Theta} \Rightarrow \theta \in \check{\Theta}) \Rightarrow z \geq y \Rightarrow \sigma_1^\omega \geq \sigma_0^\omega$. □

Lemma B.7. *For all* $\omega \in \Omega$, $\lambda \in \Theta^{[0,1]}$, *and* $\phi^\omega \in [0,1]$, $(\hat{\theta}_p + \hat{\theta}_s\tilde{h} + \hat{\theta}_s\tilde{v}(\phi^\omega) \neq \tilde{\theta}_s\tilde{v}(\phi^\omega)$ $\forall \hat{\theta}, \tilde{\theta} \in \text{supp}(\lambda)) \Rightarrow ((\hat{\sigma}_1^\omega, \hat{\sigma}_0^\omega) = (\check{\sigma}_1^\omega, \check{\sigma}_0^\omega)\, \forall \hat{\sigma}^\omega, \check{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda))$.

Proof. Consider any $\omega \in \Omega$, $\lambda \in \Theta^{[0,1]}$, and $\phi^\omega \in [0,1]$. $\hat{\theta}_p + \hat{\theta}_s\tilde{h} + \hat{\theta}_s\tilde{v}(\phi^\omega) \neq \check{\theta}_s\tilde{v}(\phi^\omega)$ $\forall \hat{\theta}, \check{\theta} \in \text{supp}(\lambda)$ implies that there is at most one $n \in \{0,1\}$ s.t. $\exists \bar{\theta} \in \text{supp}(\lambda)$ for which $n(\bar{\theta}_p + \bar{\theta}_s\tilde{h}) + \bar{\theta}_s\tilde{v}(\phi^\omega) = -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda))$.

For $1-n \in \{0,1\}$ and all $\theta \in \text{supp}(\lambda)$, $(1-n)(\theta_p+\theta_s\tilde{h})+\theta_s\tilde{v}(\phi^\omega) \neq -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda)) \Rightarrow$ $\hat{\sigma}^\omega_{1-n,\theta} = \check{\sigma}^\omega_{1-n,\theta}$ $\forall \hat{\sigma}^\omega, \check{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda) \Rightarrow \hat{\sigma}^\omega_{1-n} = \check{\sigma}^\omega_{1-n}$ $\forall \hat{\sigma}^\omega, \check{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$. For all $\check{\sigma}^\omega, \hat{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$, $\phi^\omega\check{\sigma}^\omega_1 + (1-\phi^\omega)\check{\sigma}^\omega_0 = \phi^\omega\hat{\sigma}^\omega_1 + (1-\phi^\omega)\hat{\sigma}^\omega_0$ and $\check{\sigma}^\omega_{1-n} = \hat{\sigma}^\omega_{1-n}$ implies that $\check{\sigma}^\omega_n = \hat{\sigma}^\omega_n$. Hence, $(\hat{\sigma}_1^\omega, \hat{\sigma}_0^\omega) = (\check{\sigma}_1^\omega, \check{\sigma}_0^\omega)\, \forall \hat{\sigma}^\omega, \check{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$. □

Lemma B.8. *For all* $\omega \in \Omega$, $\lambda \in \Theta^{[0,1]}$, *and* $\phi^\omega \in [0,1]$, $(\hat{n}(\hat{\theta}_p + \hat{\theta}_s\tilde{h}) + \hat{\theta}_s\tilde{v}(\phi^\omega) \neq \check{n}(\check{\theta}_p +$ $\check{\theta}_s\tilde{h}) + \check{\theta}_s\tilde{v}(\phi^\omega)$ $\forall \hat{\theta}, \check{\theta} \in \text{supp}(\lambda), \hat{n}, \check{n} \in \{0,1\}) \Rightarrow (\Sigma^{\omega*}(\phi^\omega, \lambda)$ *is a singleton*$)$.

Proof. Consider any $\omega \in \Omega$, $\lambda \in \Theta^{[0,1]}$, and $\phi^\omega \in [0,1]$. $\nexists \hat{\theta}, \check{\theta} \in \text{supp}(\lambda)$ s.t. $\hat{n}(\hat{\theta}_p + \hat{\theta}_s\tilde{h}) +$ $\hat{\theta}_s\tilde{v}(\phi^\omega) = \check{n}(\check{\theta}_p + \check{\theta}_s\tilde{h}) + \check{\theta}_s\tilde{v}(\phi^\omega)$ for some $\hat{n}, \check{n} \in \{0,1\}$ implies that there is at most one pairing of $\bar{n} \in \{0,1\}$ and $\bar{\theta} \in \text{supp}(\lambda)$ s.t. $\bar{n}(\bar{\theta}_p + \bar{\theta}_s\tilde{h}) + \bar{\theta}_s\tilde{v}(\phi^\omega) = -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda))$.

For all $\theta \in \text{supp}(\lambda)$ and $n \in \{0,1\}$, $n(\theta_p + \theta_s\tilde{h}) + \theta_s\tilde{v}(\phi^\omega) \neq -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda)) \Rightarrow$ $\sigma^\omega_{n,\theta} = \check{\sigma}^\omega_{n,\theta}$ $\forall \sigma^\omega, \check{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$. Thus, if $\nexists \bar{\theta} \in \text{supp}(\lambda)$ s.t. $\bar{n}(\bar{\theta}_p + \bar{\theta}_s\tilde{h}) + \bar{\theta}_s\tilde{v}(\phi^\omega) =$ $-\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda))$ for all $\bar{n} \in \{0,1\}$, then $\sigma^\omega = \check{\sigma}^\omega$ for all $\sigma^\omega, \check{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$.

Next, consider that $\exists!\, \bar{\theta} \in \text{supp}(\lambda)$ s.t. $\bar{n}(\bar{\theta}_p + \bar{\theta}_s\tilde{h}) + \bar{\theta}_s\tilde{v}(\phi^\omega) = -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda))$ for some $\bar{n} \in \{0,1\}$. For all $\sigma^\omega, \check{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$, $\psi^{\omega*}(\phi^\omega, \lambda) = \phi^\omega \sum_{\theta \in \text{supp}(\lambda)} \check{\sigma}^\omega_{1,\theta} + (1 -$ $\phi^\omega) \sum_{\theta \in \text{supp}(\lambda)} \check{\sigma}^\omega_{0,\theta} = \phi^\omega \sum_{\theta \in \text{supp}(\lambda)} \sigma^\omega_{1,\theta} + (1-\phi^\omega) \sum_{\theta \in \text{supp}(\lambda)} \sigma^\omega_{0,\theta}$. $\check{\sigma}^\omega_{n,\theta} = \sigma^\omega_{n,\theta}$ if $n \neq \bar{n} \vee \theta \neq \bar{\theta} \Rightarrow$ $\phi^\omega \sum_{\theta \in \text{supp}(\lambda)/\bar{\theta}} \check{\sigma}^\omega_{1,\theta} + (1-\phi^\omega) \sum_{\theta \in \text{supp}(\lambda)/\bar{\theta}} \check{\sigma}^\omega_{0,\theta} = \phi^\omega \sum_{\theta \in \text{supp}(\lambda)/\bar{\theta}} \sigma^\omega_{1,\theta} + (1-\phi^\omega) \sum_{\theta \in \text{supp}(\lambda)/\bar{\theta}} \sigma^\omega_{0,\theta} \Rightarrow$ $\phi^\omega\check{\sigma}^\omega_{1,\bar{\theta}} + (1-\phi^\omega)\check{\sigma}^\omega_{0,\bar{\theta}} = \phi^\omega\sigma^\omega_{1,\bar{\theta}} + (1-\phi^\omega)\sigma^\omega_{0,\bar{\theta}}$. $\sigma^\omega_{1-\bar{n},\theta} = \check{\sigma}^\omega_{1-\bar{n},\bar{\theta}} \Rightarrow \sigma^\omega_{\bar{n},\theta} = \check{\sigma}^\omega_{\bar{n},\bar{\theta}}$. Thus, $\sigma^\omega = \check{\sigma}^\omega$ $\forall \sigma^\omega, \check{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$. □

Lemma B.9. *For all* $\omega \in \Omega$ *and* $\lambda \in [0,1]^{|\Theta|}$, $\lim_{x \to 1}(\psi^{\omega*}(x, \lambda)) = \psi^{\omega*}(1, \lambda)$.

Proof. Lemma B.5 proves that $\psi^{\omega*}$ is non-decreasing in ϕ^ω. Thus, $\psi^\omega(\phi^\omega, \lambda) \leq \psi^\omega(1, \lambda)$ for all $\phi^\omega < 1$. It remains to be shown that $\forall \epsilon > 0$, $\exists \delta > 0$ s.t. $(\phi^\omega > 1 - \delta \Rightarrow \psi^\omega(\phi^\omega, \lambda) >$ $\psi^\omega(\phi^\omega, \lambda) - \epsilon)$.

Consider any $\epsilon > 0$. $\sum_{\theta \in \text{supp}(\lambda) \text{ s.t. } \theta_p+\theta_s\tilde{s}+\theta_s\tilde{v}(1) \geq -\Delta m^\omega(\psi^\omega(1,\lambda))} \lambda_\theta = \psi^{\omega*}(1, \lambda)$. It follows that $\sum_{\theta \in \text{supp}(\lambda) \text{ s.t. } \theta_p+\theta_s\tilde{s}+\theta_s\tilde{v}(1) > -\Delta m^\omega(\psi^\omega(1,\lambda)-\epsilon)} \lambda_\theta \geq \psi^{\omega*}(1, \lambda) > \psi^{\omega*}(1, \lambda) - \epsilon$. Let $\delta > 0$ be

s.t. for all $\phi^\omega > 1-\delta$, $\phi^\omega \sum_{\theta\in\text{supp}(\lambda) \text{ s.t. } \theta_p+\theta_s\tilde{s}+\theta_s\tilde{v}(\phi^\omega)>-\Delta m^\omega(\psi^\omega(1,\lambda)-\epsilon)} \lambda_\theta > \psi^{\omega*}(1,\lambda)-\epsilon$. Such a δ exists due to continuity of $\tilde{v}$.

At any $\phi^\omega > 1-\delta$, it must be true that $\psi^\omega(\phi^\omega,\lambda) > \psi^\omega(1,\lambda)-\epsilon$. Assume by contradiction that $\psi^\omega(\phi^\omega,\lambda) \leq \psi^\omega(1,\lambda)-\epsilon$. $\psi^\omega(\phi^\omega,\lambda) \leq \psi^\omega(1,\lambda)-\epsilon \Rightarrow -\Delta m^\omega(\psi^\omega(\phi^\omega,\lambda)) \leq -\Delta m^\omega(\psi^\omega(1,\lambda)-\epsilon) \Rightarrow (\sum_{\theta\in\text{supp}(\lambda) \text{ s.t. } \theta_p+\theta_s\tilde{s}+\theta_s\tilde{v}(\phi^\omega)>-\Delta m^\omega(\psi^\omega(1,\lambda)-\epsilon)} \lambda_\theta > \frac{\psi^{\omega*}(1,\lambda)-\epsilon}{\phi^\omega} \Rightarrow \sum_{\theta\in\text{supp}(\lambda) \text{ s.t. } \theta_p+\theta_s\tilde{s}+\theta_s\tilde{v}(\phi^\omega)>-\Delta m^\omega(\psi^\omega(\phi^\omega,\lambda))} \lambda_\theta > \frac{\psi^{\omega*}(1,\lambda)-\epsilon}{\phi^\omega}) \Rightarrow \sigma_1^{\omega*} > \frac{\psi^{\omega*}(1,\lambda)-\epsilon}{\phi^\omega}$ for all $\sigma^{\omega*} \in \Sigma^{\omega*}(\phi^\omega,\lambda) \Rightarrow \psi^{\omega*}(\phi^\omega,\lambda) \geq \sigma_1^{\omega*}\phi^\omega > \psi^{\omega*}(1,\lambda)-\epsilon$. We reached a contradiction. Therefore $\psi^\omega(\phi^\omega,\lambda) > \psi^\omega(1,\lambda)-\epsilon$ for any $\phi^\omega > 1-\delta$. Proposition B.9 is true. □

Lemma B.10. *For all* $\omega \in \Omega$ *and* $\lambda \in [0,1]^{|\Theta|}$, $\lim_{x\to 1}(\min_{\sigma^\omega\in\Sigma^{\omega*}(x,\lambda)}(\sigma_1^\omega)) = \lim_{x\to 1}(\max_{\sigma^\omega\in\Sigma^{\omega*}(x,\lambda)}(\sigma_1^\omega)) = \psi^{\omega*}(1,\lambda)$.

Proof. From $\sigma_1^\omega > \sigma_0^\omega$ for all $\sigma^\omega \in \Sigma^{\omega*}(\phi^\omega,\lambda)$ and $\psi^\omega = \phi^\omega\sigma_1^\omega + (1-\phi^\omega)\sigma_0^\omega$ follows that for all $\phi^\omega \in [0,1)$ and $\sigma^\omega \in \Sigma^{\omega*}(\phi^\omega,\lambda)$: $\min_{\sigma^\omega\in\Sigma^{\omega*}(\phi^\omega,\lambda)}(\sigma_1^\omega) \geq \psi^{\omega*}(\phi^\omega,\lambda)$ and $\max_{\sigma^\omega\in\Sigma^{\omega*}(x,\lambda)}(\sigma_1^\omega) \leq \frac{\psi^{\omega*}(\phi^\omega,\lambda)}{\phi^\omega}$. Proposition B.10 follows from $\lim_{x\to 1}(\psi^{\omega*}(x,\lambda)) = \lim_{x\to 1}(\frac{\psi^{\omega*}(x,\lambda)}{x}) = \psi^{\omega*}(1,\lambda)$. □

Lemma B.11. *Consider any* $\omega \in \Omega$, $\lambda \in [0,1]^{|\Theta|}$, $\phi^\omega \in [0,1]$, $\sigma^\omega \in \Sigma^{\omega*}(\phi^\omega,\lambda)$, *and* $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega,\hat{\lambda})$. $\forall\epsilon>0\ \exists\delta>0$ *s.t.* $\forall\hat{\lambda}\in[0,1]^{|\Theta|}$: $\sum_{\theta\in\Theta}|\lambda_\theta-\hat{\lambda}_\theta| < \delta \Rightarrow |\psi^\omega(\phi^\omega,\lambda)-\psi^\omega(\phi^\omega,\hat{\lambda})| < \epsilon$.

Proof. Consider any $\omega\in\Omega$, $\lambda,\hat{\lambda}\in[0,1]^{|\Theta|}$, $\phi^\omega\in[0,1]$, $\sigma^\omega\in\Sigma^{\omega*}(\phi^\omega,\lambda)$, $\hat{\sigma}^\omega\in\Sigma^{\omega*}(\phi^\omega,\hat{\lambda})$, and $\epsilon>0$. Let $0<\delta<\epsilon$. Suppose $\hat{\lambda}$ satisfies $\sum_{\theta\in\Theta}|\lambda_\theta-\hat{\lambda}_\theta|<\delta$.

Suppose that $\psi^{\omega*}(\phi^\omega,\lambda) > \psi^{\omega*}(\phi^\omega,\hat{\lambda})$ (analogously for $\psi^{\omega*}(\phi^\omega,\lambda) < \psi^{\omega*}(\phi^\omega,\hat{\lambda})$). For all $\theta\in\text{supp}(\lambda)$ and $n\in\{0,1\}$, $\psi^{\omega*}(\phi^\omega,\lambda) > \psi^{\omega*}(\phi^\omega,\hat{\lambda}) \Rightarrow -\Delta m^\omega(\psi^{\omega*}(\phi^\omega,\lambda)) > -\Delta m^\omega(\psi^{\omega*}(\phi^\omega,\hat{\lambda})) \Rightarrow (n(\theta_s\tilde{h}+\theta_p)+\theta_s\tilde{v}(\phi^\omega) \geq -\Delta m^\omega(\psi^{\omega*}(\phi^\omega,\lambda)) \Rightarrow n(\theta_s\tilde{h}+\theta_p)+\theta_s\tilde{v}(\phi^\omega) > -\Delta m^\omega(\psi^{\omega*}(\phi^\omega,\hat{\lambda}))) \Rightarrow \hat{\sigma}^\omega_{n,\theta} \geq \sigma^\omega_{n,\theta} \Rightarrow \psi^{\omega*}(\phi^\omega,\hat{\lambda}) \geq \phi^\omega\sum_{\theta\in\text{supp}(\lambda)}\hat{\lambda}_\theta\sigma^\omega_{1,\theta} + (1-\phi^\omega)\sum_{\theta\in\text{supp}(\lambda)}\hat{\lambda}_\theta\sigma^\omega_{0,\theta}$. Moreover, $\psi^{\omega*}(\phi^\omega,\lambda) = \phi^\omega\sum_{\theta\in\text{supp}(\lambda)}\lambda_\theta\sigma^\omega_{1,\theta} + (1-\phi^\omega)\sum_{\theta\in\text{supp}(\lambda)}\lambda_\theta\sigma^\omega_{0,\theta}$. $0 < \psi^{\omega*}(\phi^\omega,\lambda) - \psi^{\omega*}(\phi^\omega,\hat{\lambda}) \leq \phi^\omega\sum_{\theta\in\text{supp}(\lambda)}(\lambda_\theta-\hat{\lambda}_\theta)\sigma^\omega_{1,\theta} + (1-\phi^\omega)\sum_{\theta\in\text{supp}(\lambda)}(\lambda_\theta-\hat{\lambda}_\theta)\sigma^\omega_{0,\theta} \leq \sum_{\theta\in\text{supp}(\lambda)}(\lambda_\theta-\hat{\lambda}_\theta) \leq \sum_{\theta\in\Theta}(\lambda_\theta-\hat{\lambda}_\theta) < \delta < \epsilon$. Hence, $\sum_{\theta\in\Theta}|\lambda_\theta-\hat{\lambda}_\theta|<\delta$ implies that $|\psi^{\omega*}(\phi^\omega,\lambda)-\psi^{\omega*}(\phi^\omega,\hat{\lambda})|<\delta$. Thus, lemma B.11 is true. □

Lemma B.12. *Consider any* $\omega\in\Omega$, $\phi^\omega\in(0,1)$, *and* $\lambda\in[0,1]^{|\Theta|}$ *s.t.* $\forall\theta,\bar{\theta}\in\text{supp}(\lambda)$, $\bar{\theta}_s\tilde{v}(\phi^\omega) \neq \theta_s\tilde{v}(\phi^\omega)+\theta_s\tilde{h}+\theta_p$. $\forall\epsilon>0\ \exists\delta>0$ *s.t.* $\forall\hat{\lambda}\in[0,1]^{|\Theta|}$, $\sum_{\theta\in\Theta}|\lambda_\theta-\hat{\lambda}_\theta|<\delta$ *implies that for all* $\sigma^\omega\in\Sigma^{\omega*}(\phi^\omega,\lambda)$ *and* $\hat{\sigma}^\omega\in\Sigma^{\omega*}(\phi^\omega,\hat{\lambda})$, $|\sigma^\omega_n-\hat{\sigma}^\omega_n|<\epsilon\ \forall n\in\{0,1\}$.

Proof. Consider any $\omega \in \Omega$, $\phi^\omega \in (0,1)$, and λ s.t. $\forall \theta, \bar{\theta} \in \text{supp}(\lambda)$, $\bar{\theta}_s \tilde{v}(\phi^\omega) < \theta_s \tilde{v}(\phi^\omega) + \theta_s \tilde{h} + \theta_p$. Consider any $\epsilon > 0$. Let $\delta = \epsilon \times \min\{\frac{\phi^\omega}{1-\phi^\omega}, \frac{1-\phi^\omega}{\phi^\omega}\}$. Note, $\delta < \epsilon$. Let $\hat{\lambda} \in [0,1]^{|\Theta|}$ be s.t. $\sum_{\theta \in \Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \delta$. Throughout, we investigate differences in any $\sigma \in \Sigma^{\omega*}(\phi^\omega, \lambda)$ and $\hat{\sigma} \in \Sigma^{\omega*}(\phi^\omega, \hat{\lambda})$. To prove the proposition, we distinguish three cases: (1) $\psi^{\omega*}(\phi^\omega, \lambda) = \psi^{\omega*}(\phi^\omega, \hat{\lambda})$, (2) $\psi^{\omega*}(\phi^\omega, \lambda) < \psi^{\omega*}(\phi^\omega, \hat{\lambda})$, and (3) $\psi^{\omega*}(\phi^\omega, \lambda) > \psi^{\omega*}(\phi^\omega, \hat{\lambda})$.

First, we look at $\psi^{\omega*}(\phi^\omega, \lambda) = \psi^{\omega*}(\phi^\omega, \hat{\lambda})$. Let $n \in \{0,1\}$ be s.t. for all $\theta \in \text{supp}(\lambda)$, $n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi^\omega) \neq -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda))$. Such an n exists since $\bar{\theta}_s \tilde{v}(\phi^\omega) \neq \check{\theta}_s \tilde{v}(\phi^\omega) + \check{\theta}_s \tilde{h} + \check{\theta}_p$ for all $\check{\theta}, \bar{\theta} \in \text{supp}(\lambda)$. For all $\theta \in \text{supp}(\lambda)$, $((n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi^\omega) > -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda)) \Rightarrow n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi^\omega) > -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \hat{\lambda})))$ and $(n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi^\omega) < -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda)) \Rightarrow n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi^\omega) < -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \hat{\lambda})))) \Rightarrow \hat{\sigma}^\omega_{n,\theta} = \sigma^\omega_{n,\theta}$. $(\hat{\sigma}^\omega_{n,\theta} = \sigma^\omega_{n,\theta} \forall \theta \in \text{supp}(\lambda) \wedge \sigma^\omega_n = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \sigma^\omega_{n,\theta} \wedge \hat{\sigma}^\omega_n = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \hat{\sigma}^\omega_{n,\theta} + \sum_{\theta \in \Theta} (\hat{\lambda}_\theta - \lambda_\theta) \hat{\sigma}^\omega_{n,\theta}) \Rightarrow \hat{\sigma}^\omega_n = \sigma^\omega_n + \sum_{\theta \in \Theta} (\hat{\lambda}_\theta - \lambda_\theta) \hat{\sigma}^\omega_{n,\theta}$. $\sum_{\theta \in \text{supp}(\lambda) \cup \text{supp}(\hat{\lambda})} (\hat{\lambda}_\theta - \lambda_\theta) \hat{\sigma}^\omega_{n,\theta} < \delta \Rightarrow \sigma^\omega_n + \sum_{\theta \in \Theta} (\hat{\lambda}_\theta - \lambda_\theta) \hat{\sigma}^\omega_{n,\theta} < \sigma^\omega_n + \delta \Rightarrow |\sigma^\omega_n - \hat{\sigma}^\omega_n| < \epsilon$.

Moreover, let $x = |n - 1 + \phi^\omega|$. Recall $\delta < \epsilon$, $\delta \frac{x}{1-x} < \epsilon$, and $|\sigma^\omega_n - \hat{\sigma}^\omega_n| < \epsilon$. $\psi^{\omega*}(\phi^\omega, \lambda) = \psi^{\omega*}(\phi^\omega, \hat{\lambda}) \Rightarrow x\sigma^\omega_n + (1-x)\sigma^\omega_{1-n} = x\hat{\sigma}^\omega_n + (1-x)\hat{\sigma}^\omega_{1-n} \Rightarrow |\sigma^\omega_{1-n} - \hat{\sigma}^\omega_{1-n}| = |\sigma^\omega_n - \hat{\sigma}^\omega_n| \times \frac{x}{1-x} < \delta \times \frac{x}{1-x} = \epsilon \times \min\{\frac{\phi^\omega}{1-\phi^\omega}, \frac{1-\phi^\omega}{\phi^\omega}\} \times \frac{x}{1-x}$. $(\min\{\frac{\phi^\omega}{1-\phi^\omega}, \frac{1-\phi^\omega}{\phi^\omega}\} < 1$ and $(\frac{x}{1-x} = \min\{\frac{\phi^\omega}{1-\phi^\omega}, \frac{1-\phi^\omega}{\phi^\omega}\}$ or $\frac{x}{1-x} = 1/\min\{\frac{\phi^\omega}{1-\phi^\omega}, \frac{1-\phi^\omega}{\phi^\omega}\}))) \Rightarrow \frac{x}{1-x} \times \min\{\frac{\phi^\omega}{1-\phi^\omega}, \frac{1-\phi^\omega}{\phi^\omega}\} \leq 1 \Rightarrow \delta \frac{x}{1-x} < \epsilon$. Thus, $|\sigma^\omega_{1-n} - \hat{\sigma}^\omega_{1-n}| < \epsilon$.

Second, we look at $\psi^{\omega*}(\phi^\omega, \lambda) < \psi^{\omega*}(\phi^\omega, \hat{\lambda})$. For all $\theta \in \text{supp}(\lambda)$ and $n \in \{0,1\}$, $\psi^{\omega*}(\phi^\omega, \lambda) < \psi^{\omega*}(\phi^\omega, \hat{\lambda}) \Rightarrow -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda)) < -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \hat{\lambda})) \Rightarrow (n(\theta_s \tilde{h} + \theta_p) + \theta_s \tilde{v}(\phi^\omega) \leq -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \lambda)) \Rightarrow n(\theta_s \tilde{h} + \theta_p) + \theta_s \tilde{v}(\phi^\omega) < -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \hat{\lambda}))) \Rightarrow \hat{\sigma}^\omega_{n,\theta} \leq \sigma^\omega_{n,\theta} \Rightarrow \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \hat{\sigma}^\omega_{n,\theta} \leq \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \sigma^\omega_{n,\theta} = \sigma^\omega_n$. $\hat{\sigma}^\omega_n = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \hat{\sigma}^\omega_{n,\theta} + \sum_{\theta \in \Theta} (\hat{\lambda}_\theta - \lambda_\theta) \hat{\sigma}^\omega_{n,\theta}$. $(\sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \hat{\sigma}^\omega_{n,\theta} \leq \sigma^\omega_n \wedge \sum_{\theta \in \Theta} (\hat{\lambda}_\theta - \lambda_\theta) \hat{\sigma}^\omega_{n,\theta} < \delta) \Rightarrow \hat{\sigma}^\omega_n < \sigma^\omega_n + \delta$. Thus, $\hat{\sigma}^\omega_1 < \sigma^\omega_1 + \delta$ and $\hat{\sigma}^\omega_0 < \sigma^\omega_0 + \delta$.

Let $n \in \{0,1\}$ be s.t. $\hat{\sigma}^\omega_n > \sigma^\omega_n$. Such an n must exist since otherwise $\psi^{\omega*}(\phi^\omega, \lambda) < \psi^{\omega*}(\phi^\omega, \hat{\lambda})$ cannot be true. It follows that $\sigma^\omega_n < \hat{\sigma}^\omega_n < \sigma^\omega_n + \delta \Rightarrow |\sigma^\omega_n - \hat{\sigma}^\omega_n| < \delta < \epsilon$.

Let $x = |n - 1 + \phi^\omega|$. $\psi^{\omega*}(\phi^\omega, \lambda) < \psi^{\omega*}(\phi^\omega, \hat{\lambda}) \Rightarrow x\sigma^\omega_n + (1-x)\sigma^\omega_{1-n} < x\hat{\sigma}^\omega_n + (1-x)\hat{\sigma}^\omega_{1-n}$. $\hat{\sigma}^\omega_n < \sigma^\omega_n + \delta \Rightarrow x\sigma^\omega_n + (1-x)\sigma^\omega_{1-n} < x\sigma^\omega_n + x\delta + (1-x)\hat{\sigma}^\omega_{1-n} \Rightarrow \sigma^\omega_{1-n} - \frac{x}{1-x}\delta < \hat{\sigma}^\omega_{1-n} < \sigma^\omega_{1-n} + \delta$. By same reasoning as above, $\delta \frac{x}{1-x} < \epsilon$. $(\delta < \epsilon$ and $\delta \frac{x}{1-x} < \epsilon) \Rightarrow |\sigma^\omega_n - \hat{\sigma}^\omega_n| < \epsilon$. The proof of the third case is analog to that of the second case. Therefore, we refrain from writing it out.

We have shown that for any $\epsilon > 0$, $\sum_{\theta \in \Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \epsilon \times \min\{\frac{\phi^\omega}{1-\phi^\omega}, \frac{1-\phi^\omega}{\phi^\omega}\}$ implies that

$|\sigma_n^\omega - \hat{\sigma}_n^\omega| < \epsilon$. Thus, the proposition is indeed true. □

B.2 Norms

Proof of proposition 4.4

Proof. Consider any $\lambda \in [0,1]^{|\Theta|}$ and $\omega \in \Omega$. Since $\phi^\omega = 0$ is always a rest point, it remains to be shown that $\phi^\omega = 0$ is asymptotically stable. $\phi^\omega = 0$ is asymptotically stable if for all $\hat{\phi}^\omega$ close to 0 and every behavioral distribution $\sigma^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \lambda)$ that society potentially reaches at this $\hat{\phi}^\omega$, norm dynamics are negative: $\dot{\hat{\phi}}^\omega < 0$. Thus, $\phi^\omega = 0$ is asymptotically stable if $\exists \epsilon > 0$ s.t. $\hat{\phi}^\omega(1-\hat{\phi}^\omega)[(\sigma_1^\omega - \sigma_0^\omega)(\gamma v(\hat{\phi}^\omega) + \Delta m^\omega(\psi^{\omega*}(\hat{\phi}^\omega, \lambda))) - \gamma(1-\sigma_1^\omega)h + \gamma \Delta k(\hat{\phi}^\omega)] < 0$ for all $\hat{\phi}^\omega \in (0, \epsilon)$ and $\sigma^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \lambda)$. Since $\hat{\phi}^\omega(1-\hat{\phi}^\omega) > 0 \; \forall \hat{\phi}^\omega \notin \{0,1\}$ and $\operatorname{argmin}_a -\Delta m^\omega(a) = 0$, this condition is satisfied if $(\sigma_1^\omega - \sigma_0^\omega)(\gamma v(\hat{\phi}^\omega) + \Delta m^\omega(0)) - \gamma(1-\sigma_1^\omega)h + \gamma \Delta k(\hat{\phi}^\omega) < 0 \; \forall \hat{\phi}^\omega \in (0, \epsilon), \sigma^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \lambda)$. Let ϵ be sufficiently close to 0 s.t. $\gamma \Delta k(\hat{\phi}^\omega) < 0$ and $\gamma v(\hat{\phi}^\omega) + \Delta m^\omega(0) < 0 \; \forall \hat{\phi}^\omega \in (0, \epsilon)$. Such an ϵ exists due to continuity of Δk and v. Moreover, we know that $(\sigma_1^\omega - \sigma_0^\omega) \in [0,1]$ (see lemma B.6) and $\gamma(1-\sigma_1^\omega)h \geq 0$. Thus, $(\sigma_1^\omega - \sigma_0^\omega)(\gamma v(\hat{\phi}^\omega) + \Delta m^\omega(0)) - \gamma(1-\sigma_1^\omega)h + \gamma \Delta k(0) < 0 \; \forall \hat{\phi}^\omega \in (0, \epsilon)$ and $\sigma^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \lambda)$. $\phi^\omega = 0$ is a cultural equilibrium. □

Proof of proposition 4.5

Proof. Consider any $\lambda \in [0,1]^{|\Theta|}$ and $\omega \in \Omega$. $\phi^\omega = 1$ is asymptotically stable if for all $\hat{\phi}^\omega$ close to 1 and every behavioral distribution $\sigma^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \lambda)$ that society potentially reaches at this $\hat{\phi}^\omega$, norm dynamics are positive: $\dot{\hat{\phi}}^\omega > 0$. Thus, $\phi^\omega = 1$ is asymptotically stable if $\exists \epsilon > 0$ s.t. $\hat{\phi}^\omega(1-\hat{\phi}^\omega)[(\sigma_1^\omega - \sigma_0^\omega)(\gamma v(\hat{\phi}^\omega) + \Delta m^\omega(\psi^{\omega*}(\hat{\phi}^\omega, \lambda))) - \gamma(1-\sigma_1^\omega)h + \gamma \Delta k(\hat{\phi}^\omega)] > 0$ for all $\hat{\phi}^\omega \in (1-\epsilon, 1)$ and $\sigma^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \lambda)$. Since $\hat{\phi}^\omega(1-\hat{\phi}^\omega) > 0 \; \forall \hat{\phi}^\omega \notin \{0,1\}$, this condition is satisfied if $(\sigma_1^\omega - \sigma_0^\omega)(\gamma v(\hat{\phi}^\omega) + \Delta m^\omega(\psi^{\omega*}(1, \lambda))) - \gamma(1-\sigma_1^\omega)h + \gamma \Delta k(\hat{\phi}^\omega) > 0$ for all $\hat{\phi}^\omega \in (1-\epsilon, 1)$ and $\sigma^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \lambda)$.

First, suppose $\theta_s \tilde{v}(1) < -\Delta m^\omega(\psi^{\omega*}(1, \lambda))$ for all $\theta \in \operatorname{supp}(\lambda)$. $\lim_{x \to 1}(\psi^{\omega*}(x, \lambda)) = \psi^{\omega*}(1, \lambda) \Rightarrow \lim_{x \to 1}(-\Delta m^\omega(\psi^{\omega*}(x, \lambda))) = -\Delta m^\omega(\psi^{\omega*}(1, \lambda))$. Thus, for all $\hat{\phi}^\omega$ in some neighborhood of $\phi^\omega = 1$, $\theta_s \tilde{v}(\hat{\phi}^\omega) < -\Delta m^\omega(\psi^{\omega*}(\hat{\phi}^\omega, \lambda)) \Rightarrow \sigma_0^\omega = 0 \; \forall \sigma^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \lambda)$. $(\lim_{x \to 1}(\sigma_0^{\omega*}) = 0 \wedge \lim_{x \to 1}(\sigma_1^{\omega*}) = \psi^{\omega*}(1, \lambda)) \Rightarrow \lim_{x \to 1}(\min_{\sigma^\omega \in \Sigma^{\omega*}(x, \lambda)}((\sigma_1^\omega - \sigma_0^\omega)(\gamma v(x) +$

$\Delta m^{\omega}(\psi^{\omega*}(1,\lambda))) - \gamma(1-\sigma_1^{\omega*})h + \gamma\Delta k(x))) = \psi^{\omega*}(1,\lambda)(\gamma v(1) + \Delta m^{\omega}(\psi^{\omega*}(1,\lambda))) - \gamma(1 - \psi^{\omega*}(1,\lambda))h + \gamma\Delta k(1) > 0 \Rightarrow \exists \epsilon > 0$ s.t. $\forall \hat{\phi}^{\omega} \in (1-\epsilon, 1), \sigma^{\omega} \in \Sigma^{\omega*}(\hat{\phi}^{\omega}, \lambda)$, $(\sigma_1^{\omega} - \sigma_0^{\omega})(\gamma v(\hat{\phi}^{\omega}) + \Delta m^{\omega}(\psi^{\omega*}(1,\lambda))) - \gamma(1-\sigma_1^{\omega})h + \gamma\Delta k(x) > 0$.

Next, suppose $\Delta k(1) > (1 - \psi^{\omega*}(1,\lambda))h$. ($\Delta k(1) > (1 - \psi^{\omega*}(1,\lambda))h$ and $\lim_{x\to 1}(\sigma_1^{\omega*}) = \psi^{\omega*}(1,\lambda)) \Rightarrow \lim_{x\to 1}(\min_{\sigma^{\omega}\in\Sigma^{\omega*}(x,\lambda)}(\Delta k(x) - (1-\sigma_1^{\omega})h)) = \Delta k(1) - (1-\psi^{\omega*}(1,\lambda))h > 0$. Moreover, $\lim_{x\to 1}(\sigma_1^{\omega*}) = \psi^{\omega*}(1,\lambda) \Rightarrow \lim_{x\to 1}(\min_{\sigma^{\omega}\in\Sigma^{\omega*}(x,\lambda)}(\sigma_1^{\omega}(\gamma v(x) + \Delta m^{\omega}(\psi^{\omega*}(x,\lambda))) - \gamma(1-\sigma_1^{\omega})h + \gamma\Delta k(x))) = \psi^{\omega*}(1,\lambda)(\gamma v(1) + \Delta m^{\omega}(\psi^{\omega*}(1,\lambda))) - \gamma(1-\psi^{\omega*}(1,\lambda))h + \gamma\Delta k(1) > 0$. Hence, there is some $\epsilon > 0$ s.t. $\forall \hat{\phi}^{\omega} \in (1-\epsilon, 1)$ and $\sigma^{\omega} \in \Sigma^{\omega*}(\hat{\phi}^{\omega}, \lambda)$, $-\gamma(1-\sigma_1^{\omega})h + \gamma\Delta k(\hat{\phi}^{\omega}) > 0$ and $\sigma_1^{\omega}(\gamma v(\hat{\phi}^{\omega}) + \Delta m^{\omega}(\psi^{\omega*}(\hat{\phi}^{\omega}, \lambda))) - \gamma(1-\sigma_1^{\omega})h + \gamma\Delta k(\hat{\phi}^{\omega})$. Since $(\sigma_1^{\omega} - \sigma_0^{\omega}) \in [0, \sigma_1^{\omega}]$ (see lemma B.6), $(\sigma_1^{\omega} - \sigma_0^{\omega})(\gamma v(\hat{\phi}^{\omega}) + \Delta m^{\omega}(\psi^{\omega*}(\hat{\phi}^{\omega}, \lambda))) - \gamma(1-\sigma_1^{\omega})h + \gamma\Delta k(\hat{\phi}^{\omega}) > 0$.

Thus, the stated conditions imply that $\phi^{\omega} = 1$ is a cultural equilibrium at λ. □

Proof of proposition 4.6

Proof. Consider any $\omega \in \Omega$ and $\lambda \in [0,1]^{|\Theta|}$ s.t. $\phi^{\omega*} = 1$ is a cultural equilibrium satisfying proposition 4.5. First, consider the case of $\psi^{\omega*}(1,\lambda)(\gamma v(1) + \Delta m^{\omega}(\psi^{\omega*}(1,\lambda))) - \gamma(1 - \psi^{\omega*}(1,\lambda))h + \gamma\Delta k(1) > 0$ and $\Delta k(1) > (1 - \psi^{\omega*}(1,\lambda))h$. Since these inequalities are strict, there is some $\epsilon > 0$ s.t. for all $x \in (\psi^{\omega*}(1,\lambda) - \epsilon, \psi^{\omega*}(1,\lambda) + \epsilon)$, $x(\gamma v(1) + \Delta m^{\omega}(x)) - \gamma(1 - x)h + \gamma\Delta k(1) > 0$ and $\Delta k(1) > (1-x)h$. Consider any such ϵ. Lemma B.11 implies that there is a neighborhood U of λ s.t. $\hat{\lambda} \in U \Rightarrow \psi^{\omega*}(1,\hat{\lambda}) \in (\psi^{\omega*}(1,\lambda) - \epsilon, \psi^{\omega*}(1,\lambda) + \epsilon)$. Hence, there is a neighborhood U of λ s.t. $\hat{\lambda} \in U$ implies that the sufficient conditions for a perfect social norm equilibrium are satisfied at $\hat{\lambda}$. Thus, $\phi^{\omega*} = 1$ is a cultural equilibrium for all $\hat{\lambda} \in U$ for some U of λ.

Next, consider the case of $\psi^{\omega*}(1,\lambda)(\gamma v(1) + \Delta m^{\omega}(\psi^{\omega*}(1,\lambda))) - \gamma(1 - \psi^{\omega*}(1,\lambda))h + \gamma\Delta k(1) > 0$ and $\theta_s \tilde{v}(1) < -\Delta m^{\omega}(\psi^{\omega*}(1,\lambda))$ for all $\theta \in \text{supp}(\lambda)$. Recall from the proof of proposition 4.5 that $\lim_{x\to 1}(\sigma_0^{\omega*}) = 0$ and $\lim_{x\to 1}(\sigma_1^{\omega*}) = \psi^{\omega*}(1,\lambda)$ at λ. For any $\hat{\lambda} \in [0,1]^{|\Theta|}$, $\phi^{\omega} = 1$ is a cultural equilibrium if for all x in some neighborhood of 1 society reaches a behavioral distribution $\sigma^{\omega} \in \Sigma^{\omega*}(x, \hat{\lambda})$ s.t. $(\sigma_1^{\omega} - \sigma_0^{\omega})(\gamma v(x) + \Delta m^{\omega}(\psi^{\omega*}(x,\hat{\lambda}))) - \gamma(1-\sigma_1^{\omega})h + \gamma\Delta k(x) > 0$. This holds if $\lim_{x\to 1}(\min_{\sigma^{\omega}\in\Sigma^{\omega*}(x,\hat{\lambda})}((\sigma_1^{\omega} - \sigma_0^{\omega})(\gamma v(x) + \Delta m^{\omega}(\psi^{\omega*}(x,\hat{\lambda}))) - \gamma(1-\sigma_1^{\omega})h + \gamma\Delta k(x))) = \lim_{x\to 1}(\min_{\sigma^{\omega}\in\Sigma^{\omega*}(x,\hat{\lambda})}((\psi^{\omega*}(1,\hat{\lambda}) - \sigma_0^{\omega})(\gamma v(1) + \Delta m^{\omega}(\psi^{\omega*}(1,\hat{\lambda}))) - \gamma(1 - \psi^{\omega*}(1,\hat{\lambda}))h + \gamma\Delta k(1))) > 0$. There is $\alpha > 0$ s.t. for all $y_1 \in (\psi^{\omega*}(1,\lambda) - \alpha, \psi^{\omega*}(1,\lambda) + \alpha)$

and $y_0 \in [0,\alpha)$, $(y_1 - y_0)(\gamma v(1) + \Delta m^{\omega}(y_1)) - \gamma(1-y_1)h + \gamma\Delta k(1) > 0$. Consider any such α. Let $\delta \in (0,\alpha)$ be s.t. $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta}|\lambda_\theta - x_\theta| < \delta\} \Rightarrow (\psi^{\omega*}(1,\hat{\lambda}) \in (\psi^{\omega*}(1,\lambda) - \alpha, \psi^{\omega*}(1,\lambda) + \alpha)$ and $-\Delta m^{\omega}(\psi^{\omega*}(1,\hat{\lambda})) > \tilde{v}(1)\,\forall\theta \in \text{supp}(\lambda))$. Such δ exists by lemma B.11 and continuity of all involved functions. $-\Delta m^{\omega}(\psi^{\omega*}(1,\hat{\lambda})) > \tilde{v}(1)\,\forall\theta \in \text{supp}(\lambda) \Rightarrow \exists\epsilon > 0$ s.t. $-\Delta m^{\omega}(\psi^{\omega*}(x,\hat{\lambda})) > \tilde{v}(x)\,\forall x \in (1-\epsilon,1), \theta \in \text{supp}(\lambda) \Rightarrow \sigma^{\omega}_{0,\theta} = 0\,\forall\theta \in \text{supp}(\lambda), x \in (1-\epsilon,1), \sigma^{\omega} \in \Sigma^{\omega*}(x,\hat{\lambda})$. Consider any such ϵ. It follows that for all $\theta \in \text{supp}(\lambda), x \in (1-\epsilon,1)$, and $\sigma^{\omega} \in \Sigma^{\omega*}(x,\hat{\lambda})$, $\sigma^{\omega}_0 = \sum_{\theta\in\text{supp}(\hat{\lambda})}\hat{\lambda}_\theta\sigma^{\omega}_{0,\theta} = \sum_{\theta\in\Theta}(\hat{\lambda}_\theta - \lambda_\theta)\sigma^{\omega}_{0,\theta} + \sum_{\theta\in\text{supp}(\lambda)}\lambda_\theta\sigma^{\omega}_{0,\theta} = \sum_{\theta\in\Theta}(\hat{\lambda}_\theta - \lambda_\theta)\sigma^{\omega}_{0,\theta} \leq \sum_{\theta\in\Theta}(\hat{\lambda}_\theta - \lambda_\theta) \leq \delta$. Thus, $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta}|\lambda_\theta - x_\theta| < \delta\} \Rightarrow \sigma^{\omega}_0 \leq \alpha\,\forall x \in (1-\epsilon,1), \sigma^{\omega} \in \Sigma^{\omega*}(x,\hat{\lambda}) \Rightarrow \lim_{x\to1}(\max_{\sigma^{\omega}\in\Sigma^{\omega*}(x,\hat{\lambda})}(\sigma^{\omega}_0) \leq \alpha$. ($\lim_{x\to1}\max_{\sigma\in\Sigma^{\omega*}(x,\hat{\lambda})}(\sigma^{\omega}_0) < \alpha$ and $\psi^{\omega*}(1,\hat{\lambda}) \in (\psi^{\omega*}(1,\lambda)-\alpha, \psi^{\omega*}(1,\lambda)+\alpha)) \Rightarrow \lim_{x\to1}(\min_{\sigma^{\omega}\in\Sigma^{\omega*}(x,\hat{\lambda})}((\psi$ $\sigma^{\omega}_0)(\gamma v(1) + \Delta m^{\omega}(\psi^{\omega*}(1,\hat{\lambda}))) - \gamma(1-\psi^{\omega*}(1,\hat{\lambda}))h + \gamma\Delta k(1))) > 0 \Rightarrow \phi^{\omega} = 1$ is a cultural equilibrium at $\hat{\lambda}$. □

Proof of proposition 4.7:

Proof. Consider $\phi^{\omega} \in (0,1)$ that satisfies the stated conditions. $\theta_s\tilde{v}(\phi^{\omega*}) < -\Delta m^{\omega}(\phi^{\omega}) < \theta_s\tilde{v}(\phi^{\omega*}) + \theta_s\tilde{h} + \theta_p\ \forall\theta \in \text{supp}(\lambda) \Rightarrow (\sigma^{\omega}_1, \sigma^{\omega}_0) = (1,0)\,\forall\sigma^{\omega} \in \Sigma^{\omega*}(\phi^{\omega},\lambda)$. Moreover, $\exists\epsilon > 0$ s.t. $\theta_s\tilde{v}(\hat{\phi}^{\omega}) < -\Delta m^{\omega}(\hat{\phi}^{\omega}) < \theta_s\tilde{v}(\hat{\phi}^{\omega}) + \theta_s\tilde{h} + \theta_p\ \forall\theta \in \text{supp}(\lambda)$ and $\hat{\phi}^{\omega} \in (\phi^{\omega}-\epsilon, \phi^{\omega}+\epsilon) \Rightarrow (\sigma^{\omega}_1, \sigma^{\omega}_0) = (1,0)$ for all $\hat{\phi}^{\omega} \in (\phi^{\omega}-\epsilon, \phi^{\omega}+\epsilon)$ and $\sigma^{\omega} \in \Sigma^{\omega*}(\hat{\phi}^{\omega},\lambda)$. Thus, the equilibrium values $\sigma^{\omega*}_0$, $\sigma^{\omega*}_1$, and $\psi^{\omega*}(\phi^{\omega},\lambda)$ are continuous and differentiable at ϕ^{ω}; $\frac{\mathrm{d}\sigma^{\omega*}_0}{\mathrm{d}x}|_{x=\phi^{\omega}} = 0$, $\frac{\mathrm{d}\sigma^{\omega*}_1}{\mathrm{d}x}|_{x=\phi^{\omega}} = 0$, and $\frac{\partial\psi^{\omega*}(x,\lambda)}{\partial x}|_{x=\phi^{\omega}} = 1$.

Consider norm dynamics 4.2. $((\sigma^{\omega}_1, \sigma^{\omega}_0) = (1,0)$ and $\gamma(v(\phi^{\omega*}) + \Delta k(\phi^{\omega*})) = -\Delta m^{\omega}(\phi^{\omega*}))$ $\Rightarrow \dot{\phi}^{\omega} = 0$. Thus, ϕ^{ω} is a rest point. $\gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi^{\omega}}) < -\frac{\mathrm{d}\Delta m^{\omega}(\phi^{\omega})}{\mathrm{d}x}|_{x=\phi^{\omega}}$ ensures asymptotic stability. To see this, note that since equilibrium behavior satisfies $(\sigma^{\omega*}_1, \sigma^{\omega*}_0) = (1,0)$ in some interval around ϕ^{ω}, we can write norm dynamics at ϕ^{ω} as a function of the social norm only. Moreover, $\hat{\phi}^{\omega}(1-\hat{\phi}^{\omega}) > 0\ \forall\hat{\phi}^{\omega} \in (\phi^{\omega}-\epsilon, \phi^{\omega}+\epsilon)$. Thus, a rest point is asymptotically stable if

$$\frac{\mathrm{d}[C^{\omega}_1(\sigma^{\omega*},x) - C^{\omega}_0(\sigma^{\omega*},x)]}{\mathrm{d}x}|_{x=\phi^{\omega}} = \gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi^{\omega}} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi^{\omega}}) + \frac{\mathrm{d}\Delta m^{\omega}(x)}{\mathrm{d}x}|_{x=\phi^{\omega}} < 0.$$

Consequently, ϕ^{ω} is asymptotically stable under the stated conditions and, thus, a cultural equilibrium. □

Proof of proposition 4.8:

Proof. Consider any $\omega \in \Omega$, $\lambda \in [0,1]^{|\Theta|}$, $\phi^{\omega*} \in (0,1)$ satisfying proposition 4.7, and $\epsilon > 0$. Let $\check{\eta} \in (0,\epsilon)$ be s.t. for all $\theta, \bar{\theta} \in \text{supp}(\lambda)$ and $x \in [\phi^\omega - \check{\eta}, \phi^\omega + \check{\eta}]$, $\bar{\theta}_s \tilde{v}(x) < \theta_p + \theta_s \tilde{h} + \theta_s \tilde{v}(x)$. Such an $\check{\eta}$ exists since (1) $\bar{\theta}_s \tilde{v}(\phi^{\omega*}) < \theta_p + \theta_s \tilde{h} + \theta_s \tilde{v}(\phi^{\omega*})$ for all $\theta, \bar{\theta} \in \text{supp}(\lambda)$ and (2) $\tilde{v}$ is continuous. Let $\hat{\eta} \in (0,\epsilon)$ be s.t. at λ, $\dot{x} > 0$ for all $x \in [\phi^\omega - \hat{\eta}, \phi^\omega)$ and $\dot{x} < 0$ for all $x \in (\phi^\omega, \phi^\omega + \hat{\eta}]$. Such an $\hat{\eta}$ exists since ϕ^ω is a cultural equilibrium at λ. Let $\eta := \min\{\hat{\eta}, \check{\eta}\}$, $\overline{x} = \phi^\omega + \eta$, and $\underline{x} = \phi^\omega - \eta$.

$\eta \leq \hat{\eta} \Rightarrow \dot{\underline{x}} > 0$ at λ. Hence, $\eta = \min\{\hat{\eta}, \check{\eta}\} \Rightarrow (\dot{\underline{x}} > 0$ at λ and $(\check{\sigma}_1^\omega, \check{\sigma}_0^\omega) = (\hat{\sigma}_1^\omega, \hat{\sigma}_0^\omega)\, \forall \check{\sigma}^\omega, \hat{\sigma}^\omega \in \Sigma^{\omega*}(\underline{x}, \lambda)) \Rightarrow (\sigma_1^\omega - \sigma_0^\omega)(\gamma v(\underline{x}) + \Delta m^\omega(\underline{x}\sigma_1^\omega + (1-\underline{x})\sigma_0^\omega)) - \gamma(1-\underline{x})\tilde{h} + \gamma \Delta k(\underline{x}) > 0\, \forall \sigma^\omega \in \Sigma^{\omega*}(\underline{x}, \lambda)$. Consider σ_0^ω and σ_1^ω for all $\sigma^\omega \in \Sigma^{\omega*}(\underline{x}, \lambda)$. Since $-\Delta m^\omega$, $\tilde{v}$, and $\tilde{h}$ are continuous, $\exists \hat{\alpha} > 0$ s.t. $|\sigma_n^{\omega*} - \hat{\sigma}_n^\omega| < \hat{\alpha}\, \forall n \in \{0,1\} \Rightarrow (\hat{\sigma}_1^\omega - \hat{\sigma}_0^\omega)(\gamma v(\underline{x}) + \Delta m^\omega(\underline{x}\hat{\sigma}_1^\omega + (1-\underline{x})\hat{\sigma}_0^\omega)) - \gamma(1-\underline{x})\tilde{h} + \gamma \Delta k(\underline{x}) > 0$. Consider any such $\hat{\alpha} > 0$. Proposition B.12 implies that $\exists \underline{\delta}$ s.t. $\forall \hat{\lambda} \in [0,1]^{|\Theta|}$, $\sum_{\theta \in \Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \underline{\delta}$ implies that for all $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\underline{x}, \hat{\lambda})$ and $\sigma^\omega \in \Sigma^{\omega*}(\underline{x}, \lambda)$, $|\sigma_n^\omega - \hat{\sigma}_n^\omega| < \hat{\alpha}\ \forall n \in \{0,1\}$. Consequently, $\sum_{\theta \in \Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \underline{\delta} \Rightarrow \dot{\underline{x}} > 0$ at any behavior $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\underline{x}, \hat{\lambda})$. Analogously, we can show that there is some $\overline{\delta} > 0$ s.t. for all $\hat{\lambda} \in [0,1]^{|\Theta|}$, $\sum_{\theta \in \Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \overline{\delta}$ implies that $\dot{\overline{x}} < 0$ at any $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\overline{x}, \hat{\lambda})$.

Let $\delta := \min\{\underline{\delta}, \overline{\delta}\}$. $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta \in \Theta} |\lambda_\theta - x_\theta| < \delta\} \Rightarrow (\dot{\underline{x}} > 0$ at any $\sigma^\omega \in \Sigma^{\omega*}(\underline{x}, \hat{\lambda})$ and $\dot{\overline{x}} < 0$ at any $\sigma^\omega \in \Sigma^{\omega*}(\overline{x}, \hat{\lambda}))$. Whenever society is at social norm $x \in \{\underline{x}, \overline{x}\}$ and preference distribution $\hat{\lambda}$, it coordinates into the behavioral equilibrium $\Sigma^{\omega*}(x, \hat{\lambda})$. At each possible Nash equilibrium $\hat{\sigma}^\omega \in \Sigma^{\omega*}(x, \hat{\lambda})$ that society reaches, the social norm x decreases if $x = \overline{x}$ and increases if $x = \overline{x}$. Hence, the social norm must evolve to some minimal asymptotically stable set $\hat{\Phi}^{\omega*} \subset (\underline{x}, \overline{x}) = (\phi^{\omega*} - \eta, \phi^{\omega*} + \eta) \subset (\phi^{\omega*} - \epsilon, \phi^{\omega*} + \epsilon)$ at $\hat{\lambda}$. □

Proposition B.1. *Consider any $\omega \in \Omega$, $\lambda \in [0,1]^{|\Theta|}$, and $\phi^{\omega*} \in (0,1)$ satisfying proposition 4.7. Moreover, let $\tau \in \mathbb{R}_{\geq 0}$ be s.t. $\tau v(\phi^{\omega*}) < -\Delta m^\omega(\phi^{\omega*}) < \tau h + \tau v(\phi^{\omega*})$. There is $\delta > 0$ and $\eta > 0$ s.t. $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta \in \Theta} |\lambda_\theta - x_\theta| < \delta\}$ implies that:*

1. *for all $\phi^\omega \in (\phi^{\omega*} - \eta, \phi^{\omega*} + \eta)$,*
   - *(a) $\tau v(\phi^\omega) < -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \hat{\lambda})) < \tau h + \tau v(\phi^\omega)$ and*
   - *(b) $\theta_s \tilde{v}(\phi^\omega) < -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \hat{\lambda})) < \theta_s \tilde{v}(\phi^\omega) + \theta_s \tilde{h} + \theta_p\, \forall \theta \in \text{supp}(\lambda)$, and*

2. *there is some* $\hat{\Phi}^\omega \subset (\phi^{\omega*} - \eta, \phi^{\omega*} + \eta)$ *s.t.*

   (a) $\hat{\Phi}^\omega$ *is a minimal asymptotically stable set at* $\hat{\lambda}$ *and*

   (b) *for all* $\check{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta}|\lambda_\theta - x_\theta| < \delta\}$*, there is some minimal asymptotically stable set* $\check{\Phi}^\omega \subset (\phi^{\omega*} - \eta, \phi^{\omega*} + \eta)$ *at* $\check{\lambda}$ *s.t. each* $\hat{\phi}^\omega \in \hat{\Phi}^\omega$ *is in it's basin of attraction.*

Proof. Consider any $\omega \in \Omega$, $\lambda \in [0,1]^{|\Theta|}$, $\phi^{\omega*} \in (0,1)$ satisfying proposition 4.7, and $\tau \in \mathbb{R}_{\geq 0}$ s.t. $\tau v(\phi^{\omega*}) < -\Delta m^\omega(\phi^{\omega*}) < \tau h + \tau v(\phi^{\omega*})$. Let $\epsilon > 0$ be s.t. for all $x \in (\phi^{\omega*} - \epsilon, \phi^{\omega*} + \epsilon)$, $\tau v(x) < -\Delta m^\omega(x) < \tau h + \tau v(x)$ and $\theta_s\tilde{v}(x) < -\Delta m^\omega(x) < \theta_s\tilde{v}(x) + \theta_s\tilde{h} + \theta_p$ $\forall \theta \in \text{supp}(\lambda)$. Such an ϵ exists since (1) $\tau v(\phi^{\omega*}) < -\Delta m^\omega(\phi^{\omega*}) < \tau h + \tau v(\phi^{\omega*})$ and $\theta_s\tilde{v}(\phi^{\omega*}) < -\Delta m^\omega(\phi^{\omega*}) < \theta_s\tilde{v}(\phi^{\omega*}) + \theta_s\tilde{h} + \theta_p \forall \theta \in \text{supp}(\lambda)$ and (2) v, $\tilde{v}$, $-\Delta m^\omega$, and Δk are continuous. For all $x \in (\phi^{\omega*} - \epsilon, \phi^{\omega*} + \epsilon)$, $\theta_s\tilde{v}(x) < -\Delta m^\omega(x) < \theta_s\tilde{v}(x) + \theta_s\tilde{h} + \theta_p \forall \theta \in \text{supp}(\lambda) \Rightarrow \psi^{\omega*}(x,\lambda) = x$. Let $\check{\alpha} > 0$ be s.t. for all $x \in (\phi^{\omega*} - \epsilon, \phi^{\omega*} + \epsilon)$ and $y \in (x - \check{\alpha}, x + \check{\alpha})$, $\tau v(x) < -\Delta m^\omega(y) < \tau h + \tau v(x)$ and $\theta_s\tilde{v}(x) < -\Delta m^\omega(y) < \theta_s\tilde{v}(x) + \theta_s\tilde{h} + \theta_p \forall \theta \in \text{supp}(\lambda)$. Similar to above, such $\check{\alpha}$ exists since $-\Delta m^\omega$ is continuous. Let $\check{\delta} > 0$ be s.t. $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta}|\lambda_\theta - x_\theta| < \check{\delta}\} \Rightarrow \psi^{\omega*}(x,\hat{\lambda}) \in (x - \check{\alpha}, x + \check{\alpha})$. Such $\check{\delta}$ exists due to proposition B.11 and $\psi^{\omega*}(x,\lambda) = x$ for all $x \in (\phi^{\omega*} - \epsilon, \phi^{\omega*} + \epsilon)$.

Let $\check{\eta} \in (0,\epsilon)$ be s.t. for all $\theta, \bar{\theta} \in \text{supp}(\lambda)$ and $x \in [\phi^\omega - \check{\eta}, \phi^\omega + \check{\eta}]$, $\bar{\theta}_s\tilde{v}(x) < \theta_p + \theta_s\tilde{h} + \theta_s\tilde{v}(x)$. Such an $\check{\eta}$ exists since (1) $\bar{\theta}_s\tilde{v}(\phi^{\omega*}) < \theta_p + \theta_s\tilde{h} + \theta_s\tilde{v}(\phi^{\omega*})$ for all $\theta, \bar{\theta} \in \text{supp}(\lambda)$ and (2) $\tilde{v}$ is continuous. Let $\hat{\eta} \in (0,\epsilon)$ be s.t. at λ, $\dot{x} > 0$ for all $x \in [\phi^\omega - \hat{\eta}, \phi^\omega)$ and $\dot{x} < 0$ for all $x \in (\phi^\omega, \phi^\omega + \hat{\eta}]$. Such an $\hat{\eta}$ exists since ϕ^ω is a cultural equilibrium at λ. Let $\eta := \min\{\hat{\eta}, \check{\eta}\}$, $\overline{x} = \phi^\omega + \eta$, and $\underline{x} = \phi^\omega - \eta$.

$\eta \leq \check{\eta} \Rightarrow \bar{\theta}_s\tilde{v}(\underline{x}) < \theta_p + \theta_s\tilde{h} + \theta_s\tilde{v}(\underline{x}) \forall \theta, \bar{\theta} \in \text{supp}(\lambda) \Rightarrow (\check{\sigma}_1^\omega, \check{\sigma}_0^\omega) = (\hat{\sigma}_1^\omega, \hat{\sigma}_0^\omega) \forall \check{\sigma}^\omega, \hat{\sigma}^\omega \in \Sigma^{\omega*}(\underline{x}, \lambda)$ (see lemma B.7). $\eta \leq \hat{\eta} \Rightarrow \dot{\underline{x}} > 0$. Hence, $\eta = \min\{\hat{\eta}, \check{\eta}\} \Rightarrow (\dot{\underline{x}} > 0 \wedge (\check{\sigma}_1^\omega, \check{\sigma}_0^\omega) = (\hat{\sigma}_1^\omega, \hat{\sigma}_0^\omega) \forall \check{\sigma}^\omega, \hat{\sigma}^\omega \in \Sigma^{\omega*}(\underline{x}, \lambda)) \Rightarrow (\sigma_1^\omega - \sigma_0^\omega)(\gamma v(\underline{x}) + \Delta m^\omega(\underline{x}\sigma_1^\omega + (1 - \underline{x})\sigma_0^\omega)) - \gamma(1 - \underline{x})\tilde{h} + \gamma\Delta k(\underline{x}) > 0 \forall \sigma^\omega \in \Sigma^{\omega*}(\underline{x}, \lambda)$. Consider σ_0^ω and σ_1^ω for all $\sigma^\omega \in \Sigma^{\omega*}(\underline{x}, \lambda)$. Since $-\Delta m^\omega$, $\tilde{v}$, and $\tilde{h}$ are continuous, $\exists \hat{\alpha} > 0$ s.t. $|\sigma_n^{\omega*} - \hat{\sigma}_n^\omega| < \hat{\alpha} \forall n \in \{0,1\} \Rightarrow (\hat{\sigma}_1^\omega - \hat{\sigma}_0^\omega)(\gamma v(\underline{x}) + \Delta m^\omega(\underline{x}\hat{\sigma}_1^\omega + (1 - \underline{x})\hat{\sigma}_0^\omega)) - \gamma(1 - \underline{x})\tilde{h} + \gamma\Delta k(\underline{x}) > 0$. Consider any such $\hat{\alpha} > 0$. Proposition B.12 implies that $\exists \underline{\delta}$ s.t. $\forall \hat{\lambda} \in [0,1]^{|\Theta|}$, $\sum_{\theta\in\Theta}|\lambda_\theta - \hat{\lambda}_\theta| < \underline{\delta}$ implies that for all $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\underline{x}, \hat{\lambda})$ and $\sigma^\omega \in \Sigma^{\omega*}(\underline{x}, \lambda)$, $|\sigma_n^\omega - \hat{\sigma}_n^\omega| < \hat{\alpha}$ $\forall n \in \{0,1\}$. Consequently, $\sum_{\theta\in\Theta}|\lambda_\theta - \hat{\lambda}_\theta| < \underline{\delta} \Rightarrow \dot{\underline{x}} > 0$ at any $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\underline{x}, \hat{\lambda})$. Analogously, we can show that there is some $\overline{\delta} > 0$ s.t. for all

$\hat{\lambda} \in [0,1]^{|\Theta|}$, $\sum_{\theta\in\Theta}|\lambda_\theta - \hat{\lambda}_\theta| < \overline{\delta}$ implies that $\dot{\overline{x}} < 0$ at any $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\overline{x}, \hat{\lambda})$.

Let $\delta := \min\{\underline{\delta}, \overline{\delta}, \check{\delta}\}$. $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta}|\lambda_\theta - x_\theta| < \delta\} \Rightarrow (\dot{\underline{x}} > 0$ at any $\sigma^\omega \in \Sigma^{\omega*}(\underline{x}, \hat{\lambda})$ and $\dot{\overline{x}} < 0$ at any $\sigma^\omega \in \Sigma^{\omega*}(\overline{x}, \hat{\lambda}))$. Whenever society is at social norm $x \in \{\underline{x}, \overline{x}\}$ and preference distribution $\hat{\lambda}$, it coordinates into the behavioral equilibrium $\Sigma^{\omega*}(x, \hat{\lambda})$. At each possible Nash equilibrium $\hat{\sigma}^\omega \in \Sigma^{\omega*}(x, \hat{\lambda})$ that society reaches, the social norm x decreases if $x = \overline{x}$ and increases if $x = \overline{x}$. Hence, the social norm must evolve to some minimal asymptotically stable set $\hat{\Phi}^{\omega*} \subset (\underline{x}, \overline{x}) \subset (\phi^{\omega*} - \eta, \phi^{\omega*} + \eta)$ at $\hat{\lambda}$. Next, we investigate norm evolution at any other $\check{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta}|\lambda_\theta - x_\theta| < \tilde{\delta}\}$, when starting at some element $\hat{\phi}^\omega \in \hat{\Phi}^\omega$. $\hat{\phi}^\omega \in \hat{\Phi}^\omega \subset (\underline{x}, \overline{x})$ and $\dot{\underline{x}} > 0 \wedge \dot{\overline{x}} < 0$ at $\check{\lambda}$ imply that there is some minimal asymptotically stable set $\check{\Phi}^\omega \subset (\underline{x}, \overline{x})$ that norms evolve to when starting at $\hat{\phi}^\omega \in \hat{\Phi}^\omega$.

Lastly, note that $\delta \leq \check{\delta}$ and $\eta < \epsilon$ imply that for all $\phi^\omega \in (\phi^{\omega*} - \eta, \phi^{\omega*} + \eta)$ and $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta}|\lambda_\theta - x_\theta| < \delta\}$, $\tau v(\phi^\omega) < -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \hat{\lambda})) < \tau h + \tau v(\phi^\omega)$ and $\theta_s\tilde{v}(\phi^\omega) < -\Delta m^\omega(\psi^{\omega*}(\phi^\omega, \hat{\lambda})) < \theta_p + \theta_s\tilde{h} + \theta_s\tilde{v}(\phi^\omega)\,\forall\theta \in \text{supp}(\lambda)$. Thus, proposition B.1 is true for $\delta = \min\{\underline{\delta}, \overline{\delta}, \check{\delta}\}$ and $\eta = \min\{\hat{\eta}, \check{\eta}\}$. □

Proof of proposition 4.9

Proof. Consider any $\omega \in \Omega$, $\lambda \in [0,1]^{|\Theta|}$, and $\phi^{\omega*} \in (0,1)$, for which proposition 4.7 holds. Moreover, consider $\hat{\lambda} \in [0,1]^{|\Theta|}$ s.t.

1. $\theta_s\tilde{v}(\phi^{\omega*}) \leq -\Delta m^\omega(\phi^{\omega*}) < \theta_s\tilde{v}(\phi^{\omega*}) + \theta_s\tilde{h} + \theta_p$ for all $\theta \in \text{supp}(\hat{\lambda})$,
2. $\theta_s\tilde{v}(\phi^{\omega*}) = -\Delta m^\omega(\phi^{\omega*})$ for some $\theta \in \text{supp}(\hat{\lambda})$, and
3. $\gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}}) > \frac{-\Delta m^\omega(\phi^{\omega*})}{\tilde{v}(\phi^{\omega*})} \times \frac{\mathrm{d}\tilde{v}(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}}$.

Since $\phi^{\omega*}$ is a cultural equilibrium at λ,

- $-\Delta m^\omega(\phi^{\omega*}) = \gamma v(\phi^{\omega*}) + \gamma\Delta k(\phi^{\omega*})$ and
- $-\frac{\mathrm{d}\Delta m^\omega(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}} > \gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}}) > \frac{-\Delta m^\omega(\phi^{\omega*})}{\tilde{v}(\phi^{\omega*})} \times \frac{\mathrm{d}\tilde{v}(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}}$.

Let $\bar{\theta}_s := \frac{-\Delta m^\omega(\phi^{\omega*})}{\tilde{v}(\phi^{\omega*})}$. Note, there is $\theta \in \text{supp}(\hat{\lambda})$ s.t. $\theta_s = \bar{\theta}_s$. In slight abuse of notation, we write $\lambda_{\bar{\theta}} := \sum_{\theta\in\text{supp}(\hat{\lambda}) \text{ s.t. } \theta_s=\bar{\theta}_s} \lambda_\theta$.

Consider some $\epsilon > 0$ s.t. for all $\theta \in \text{supp}(\hat{\lambda})$ and $x \in (\phi^{\omega*} - \epsilon, \phi^{\omega*})$, (1) $\min\{\bar{\theta}_s\tilde{v}(x), -\Delta m^\omega(x)\} < \theta_s\tilde{v}(\phi^{\omega*}) + \theta_s\tilde{h} + \theta_p$, (2) $-\Delta m^\omega(x) < \bar{\theta}_s\tilde{v}(x) < -\Delta m^\omega(x + (1-x)\lambda_{\bar{\theta}})$, and (3) $\gamma v(x) +$

$\gamma\Delta k(x) < \bar{\theta}_s\tilde{v}(x)$. Such ϵ exists since $\tilde{v}$ and $-\Delta m^\omega$ are continuous and $-\frac{\mathrm{d}\Delta m^\omega(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}} > \gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}}) > \bar{\theta}_s \times \frac{\mathrm{d}\tilde{v}(x)}{\mathrm{d}x}|_{x=\phi^{\omega*}}$. For all $x \in (\phi^{\omega*}-\epsilon, \phi^{\omega*})$, $\min\{\bar{\theta}_s\tilde{v}(x), -\Delta m^\omega(x)\} < \theta_s\tilde{v}(\phi^{\omega*}) + \theta_s\tilde{h} + \theta_p \Rightarrow \sigma_1^\omega = 1\ \forall\sigma^\omega \in \Sigma^{\omega*}(x, \hat{\lambda})$. Moreover, for all $x \in (\phi^{\omega*} - \epsilon, \phi^{\omega*})$, $-\Delta m^\omega(x) < \bar{\theta}_s\tilde{v}(x) < -\Delta m^\omega(x + (1-x)\lambda_{\bar{\theta}}) \Rightarrow -\Delta m^\omega(\psi^{\omega*}(x, \hat{\lambda})) = \bar{\theta}_s\tilde{v}(x)$. Hence, $\gamma v(x) + \gamma\Delta k(x) < \bar{\theta}_s\tilde{v}(x) = -\Delta m^\omega(\psi^{\omega*}(x, \hat{\lambda}))$. Moreover, $x < \phi^{\omega*} < \frac{1}{2} \Rightarrow \Delta k(x) < 0$.

It follows that for al $\sigma^\omega \in \Sigma^{\omega*}(x, \hat{\lambda})$, $(\sigma_1^\omega - \sigma_0^\omega)(\gamma v(x) + \Delta m^\omega(\psi^{\omega*}(x, \hat{\lambda}))) + \Delta k(x) < 0 \Rightarrow \dot{x} < 0$. Thus, norm evolution moves away from $\phi^{\omega*}$. Since norms evolve on an interval, $\phi^{\omega*}$ cannot be part of an asymptotically stable set at $\hat{\lambda}$. □

B.3 Preferences

Lemma B.13. *For all $\omega \in \Omega, \lambda \in \{x \in [0,1]^{|\Theta|} : \theta^d \in \mathrm{supp}(x)\}, \theta \in \mathrm{supp}(\lambda), \phi \in [0,1]^{|\Omega|}, n \in \{0,1\}$, and $\sigma \in \prod_{\omega\in\Omega}\Sigma^{\omega*}(\phi^\omega, \lambda)$, $B^\omega_{n,\theta^d}(\sigma^\omega, \phi^\omega) \geq B^\omega_{n,\theta}(\sigma^\omega, \phi^\omega), B^\omega_{\theta^d}(\sigma^\omega, \phi^\omega) \geq B^\omega_\theta(\sigma^\omega, \phi^\omega)$, and $B_{\theta^d}(\sigma, \phi) \geq B_\theta(\sigma, \phi)$.*

Proof. Consider any $\omega \in \Omega, \lambda \in \{x \in [0,1]^{|\Theta|} : \theta^d \in \mathrm{supp}(x)\}, \theta \in \mathrm{supp}(\lambda), \phi \in [0,1]^{|\Omega|}, n \in \{0,1\}$, and $\sigma \in \prod_{\omega\in\Omega}\Sigma^{\omega*}(\phi^\omega, \lambda)$. Since $B_{\bar{\theta}}(\sigma, \phi) = \sum_{\omega\in\Omega} B^\omega_{\bar{\theta}}(\sigma^\omega, \phi^\omega) = \sum_{\omega\in\Omega} \phi^\omega B^\omega_{1,\bar{\theta}}(\sigma^\omega, \phi^\omega) + (1-\phi^\omega)B^\omega_{0,\bar{\theta}}(\sigma^\omega, \phi^\omega)\ \forall\bar{\theta} \in \mathrm{supp}(\lambda)$, it is sufficient to show that $B^\omega_{n,\theta^d} \geq B^\omega_{n,\theta}$ for all $\omega \in \Omega, \theta \in \mathrm{supp}(\lambda), \phi^\omega \in [0,1], n \in \{0,1\}, \sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$. Assume by contradiction that $\exists\omega \in \Omega, \bar{\theta} \in \mathrm{supp}(\lambda), n \in \{0,1\}, \phi^\omega \in [0,1], \sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$ s.t. $B^\omega_{n,\theta^d}(\sigma^\omega, \phi^\omega) < B^\omega_{\bar{\theta},n}(\sigma^\omega, \phi^\omega)$. $B^\omega_{n,\theta^d}(\sigma^\omega, \phi^\omega) \neq B^\omega_{\bar{\theta},n}(\sigma^\omega, \phi^\omega)$ only if $b^\omega(a, n, \psi^{\omega*}(\phi^\omega, \lambda), \phi^\omega) > b^\omega(1-a, n, \psi^{\omega*}(\phi^\omega, \lambda), \phi^\omega) \wedge \sigma^\omega_{n,\theta^d} \neq \sigma^\omega_{n,\bar{\theta}}$. $b^\omega(a, n, \psi^{\omega*}(\phi^\omega, \lambda), \phi^\omega) > b^\omega(1-a, n, \psi^{\omega*}(\phi^\omega, \lambda), \phi^\omega) \Rightarrow u^\omega(a, n, \psi^{\omega*}(\phi^\omega, \lambda), \phi^\omega, \theta^d) > u^\omega(1-a, n, \psi^{\omega*}(\phi^\omega, \lambda), \phi^\omega, \theta^d) \Rightarrow \sigma^\omega_{n,\theta^d} = a \Rightarrow \sigma^\omega_{n,\bar{\theta}} \neq a$. However, $b^\omega(a, n, \psi^{\omega*}(\phi^\omega, \lambda), \phi^\omega) > b^\omega(1-a, n, \psi^{\omega*}(\phi^\omega, \lambda), \phi^\omega) \wedge \sigma^\omega_{n,\theta^d} = a \neq \sigma^{\omega*}_{n,\bar{\theta}} \Rightarrow B^\omega_{n,\theta^d}(\sigma^\omega, \phi^\omega) > B^\omega_{\bar{\theta},n}(\sigma^\omega, \phi^\omega)$. We have reached a contradiction. □

Lemma B.14. *For all $\lambda \in [0,1]^{|\Theta|}$ and $\phi \in [0,1]^{|\Omega|}$, $(\sigma^\omega_n = \bar{\sigma}^\omega_n\ \forall\omega \in \Omega, n \in \{0,1\}/\{1-\phi^\omega\}, \sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda^d)$, and $\bar{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)) \Rightarrow (\dot{\lambda}_\theta = 0\ \forall\theta \in \mathrm{supp}(\lambda))$.*

Proof. Consider any $\omega \in \Omega$ and $\lambda \in [0,1]^{|\Theta|}$. Note that $(\sigma^\omega_1, \sigma^\omega_0) = (\bar{\sigma}^\omega_1, \bar{\sigma}^\omega_0) \Rightarrow \psi^{\omega*}(\phi^{\omega*}, \lambda^d) = \psi^{\omega*}(\phi^{\omega*}, \lambda)$. Throughout, we write $\psi^{\omega*} := \psi^{\omega*}(\phi^{\omega*}, \lambda) = \psi^{\omega*}(\phi^{\omega*}, \lambda^d)$.

Consider any $\omega \in \Omega$ and $n \in \{0,1\}$. First, we investigate the case where $\exists a \in \{0,1\}$ s.t. $b^\omega(a, n, \psi^{\omega*}, \phi^\omega) > b^\omega(1-a, n, \psi^{\omega*}, \phi^\omega)$. For all $\theta \in \mathrm{supp}(\lambda)$, $\sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda^d)$,

and $\bar{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$, $b^\omega(a, n, \psi^{\omega*}, \phi^\omega) > b^\omega(1-a, n, \psi^{\omega*}, \phi^\omega) \Rightarrow \sigma_n^\omega = a = \bar{\sigma}_n^\omega$. $\bar{\sigma}_n^\omega = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \bar{\sigma}_{n,\theta}^\omega = a \in \{0,1\} \Rightarrow \bar{\sigma}_{n,\theta}^\omega = a \, \forall \theta \in \text{supp}(\lambda) \Rightarrow B_{\hat{\theta},n}^\omega = b^\omega(a, n, \psi^{\omega*}, \phi^{\omega*}) = B_{\tilde{\theta},n}^\omega \ \forall \tilde{\theta}, \hat{\theta} \in \text{supp}(\lambda)$. Next, we look at the case of $b^\omega(0, n, \psi^{\omega*}, \phi^\omega) = b^\omega(1, n, \psi^{\omega*}, \phi^\omega)$. $b^\omega(0, n, \psi^{\omega*}, \phi^\omega) = b^\omega(1, n, \psi^{\omega*}, \phi^\omega) \Rightarrow (1-y)b^\omega(0, n, \psi^{\omega*}, \phi^\omega) + y b^\omega(1, n, \psi^{\omega*}, \phi^\omega) = (1-x)b^\omega(0, n, \psi^{\omega*}, \phi^\omega) + x b^\omega(1, n, \psi^{\omega*}, \phi^\omega) \, \forall x, y \in [0,1] \Rightarrow B_{\hat{\theta},n}^\omega = B_{\tilde{\theta},n}^\omega \forall \tilde{\theta}, \hat{\theta} \in \text{supp}(\lambda)$.

Thus, for all $\omega \in \Omega, \lambda \in [0,1]^{|\Theta|}, \phi \in [0,1]^{|\Omega|}, n \in \{0,1\}, \sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda^d)$, and $\bar{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)$, $(\sigma_n^\omega = \bar{\sigma}_n^\omega \Rightarrow B_{\hat{\theta},n}^\omega(\bar{\sigma}^\omega, \phi^\omega) = B_{\tilde{\theta},n}^\omega(\bar{\sigma}^\omega, \phi^\omega)$. Hence, $(\sigma_n^\omega = \bar{\sigma}_n^\omega \ \forall \omega \in \Omega, \lambda \in [0,1]^{|\Theta|}, \phi \in [0,1]^{|\Omega|}, n \in \{0,1\}, \sigma^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda^d)$, and $\bar{\sigma}^\omega \in \Sigma^{\omega*}(\phi^\omega, \lambda)) \Rightarrow (\dot{\lambda}_\theta = 0 \quad \forall \theta \in \text{supp}(\lambda))$. □

Proof of lemma 5.1:

Proof. Consider any $\phi \in [0,1]^{|\Theta|}$ and $\lambda \in \Lambda_p(\phi)$. First, we look at condition 1. Assume by contradiction that there is $\bar{\theta} \in \text{supp}(\lambda)$ s.t. $\bar{\theta}_s > \min\{\frac{-\Delta m^\omega(\phi^\omega)}{\tilde{v}(\phi^\omega)} : \phi^\omega = \psi^{\omega*}(\phi^\omega, \lambda^d) \in (0,1)\}$. Throughout, let $\bar{\omega} = \text{argmin}\{\frac{-\Delta m^\omega(\phi^\omega)}{\tilde{v}(\phi^\omega)} : \phi^\omega = \psi^{\omega*}(\phi^\omega, \lambda^d) \in (0,1)\}$. $\bar{\theta}_s > \frac{-\Delta m^{\bar{\omega}}(\phi^{\bar{\omega}})}{\tilde{v}(\phi^{\bar{\omega}})} \Rightarrow \bar{\theta}_s \tilde{v}(\phi^{\bar{\omega}}) > -\Delta m^{\bar{\omega}}(\phi^{\bar{\omega}}) \Rightarrow \bar{\sigma}_0^{\bar{\omega}} > 0 \, \forall \bar{\sigma}^{\bar{\omega}} \in \Sigma^{\bar{\omega}*}(\phi^{\bar{\omega}}, \lambda)$. To see this last part, suppose $\bar{\sigma}_0^{\bar{\omega}} = 0$ for some $\bar{\sigma}^{\bar{\omega}} \in \Sigma^{\bar{\omega}*}(\phi^{\bar{\omega}}, \lambda)$. $\bar{\sigma}_0^{\bar{\omega}} = 0 \Rightarrow \psi^{\bar{\omega}*}(\phi^{\bar{\omega}}, \lambda) \leq \phi^{\bar{\omega}}$. $(\psi^{\bar{\omega}*}(\phi^{\bar{\omega}}, \lambda) \leq \phi^{\bar{\omega}} \wedge \bar{\theta}_s \tilde{v}(\phi^{\bar{\omega}}) > -\Delta m^{\bar{\omega}}(\phi^{\bar{\omega}}) \Rightarrow \bar{\sigma}_{\bar{\theta},0}^{\bar{\omega}} > 0 \, \forall \bar{\sigma}^{\bar{\omega}} \in \Sigma^{\bar{\omega}*}(\phi^{\bar{\omega}}, \lambda) \Rightarrow \bar{\sigma}_0^{\bar{\omega}} > 0 \, \forall \bar{\sigma}^{\bar{\omega}} \in \Sigma^{\bar{\omega}*}(\phi^{\bar{\omega}}, \lambda)$, which cannot be true. Hence, $\bar{\sigma}_0^{\bar{\omega}} > 0 \, \forall \bar{\sigma}^{\bar{\omega}} \in \Sigma^{\bar{\omega}*}(\phi^{\bar{\omega}}, \lambda)$. Next, note that $\psi^{\bar{\omega}*}(\phi^{\bar{\omega}}, \lambda^d)$ and $\text{supp}(\lambda^d) = \{\theta^d\}$ imply that $(\sigma_1^{\bar{\omega}}, \sigma_0^{\bar{\omega}}) = (1,0) \, \forall \sigma^{\bar{\omega}} \in \Sigma^{\bar{\omega}*}(\phi^{\bar{\omega}}, \lambda^d)$. Hence, $\lambda \notin \Lambda_p(\phi)$. We reached a contradiction implying that condition 1 must be true.

Next, lets look at condition 2. Let $\check{\omega} = \text{argmax}_{\omega \in \Omega}\{-\Delta m^\omega(1) : \phi^\omega = \psi^{\omega*}(1, \lambda^d) = 1\}$. For all $\lambda \in \Lambda_p(\phi)$, $\psi^{\check{\omega}*}(\phi^{\check{\omega}}, \lambda) = \psi^{\check{\omega}*}(\phi^{\check{\omega}}, \lambda^d) = 1 \Rightarrow \psi^{\check{\omega}*}(\phi^{\check{\omega}}, \lambda) = 1$. Hence, for all $\theta \in \text{supp}(\lambda)$, $\theta_s \tilde{v}(1) + \theta_s \tilde{h} + \theta_p \geq -\Delta m^{\check{\omega}}(1) \Rightarrow \theta_s \geq \frac{-\Delta m^{\check{\omega}}(1) - \theta_p}{\tilde{v}(1) + \tilde{h}} \Rightarrow \theta_s \tilde{v}(x) + \theta_s \tilde{h} + \theta_p \geq \frac{(-\Delta m^{\check{\omega}}(1) - \theta_p)(\tilde{v}(x) + \tilde{h})}{\tilde{v}(1) + \tilde{h}} + \theta_p \geq \frac{-\Delta m^{\check{\omega}}(1)}{\tilde{v}(1) + \tilde{h}}(\tilde{v}(x) + \tilde{h})$. Condition 2 follows straight-away. □

Proposition B.2. *If the set of situations Ω is sufficiently diverse regarding the contribution costs $\{-\Delta^\omega\}_{\omega \in \Omega}$, then there exists a ϕ_r^* of definition 5.6 that features all three types of cultural equilibria presented in section 4.2.1 at preference distribution λ^d.*

Proof. Recall that for any situation ω a cultural equilibrium of no social norm $\phi^{\omega*} = 0$ always exists. We continue by showing that (1) for some situation $\check{\omega}$ there exists the costs

of contribution $-\Delta m^{\breve{\omega}}$ s.t. a perfect social norm equilibrium of proposition 4.2 that satisfies condition 4 of definition 5.6 and (2) given the perfect social norm in situation $\breve{\omega}$, for some other situation $\bar{\omega}$ there exists the costs of contribution $-\Delta m^{\omega}$ s.t. an imperfect social norm equilibrium of proposition 4.3 that satisfies conditions 2 and 3 of definition 5.6.

We start with the former. Therefore, let there be some situation $\breve{\omega}$ for which the costs of contribution satisfy $\theta_s^d\tilde{v}(1) < -\Delta m^{\breve{\omega}}(1) < \min\{\theta_s^d\tilde{v}(1) + \theta_s^d\tilde{h} + \theta_p^d, \gamma v(1) + \gamma h\}$. Hence, the equilibrium cooperation share at the perfect social norm $\phi^{\breve{\omega}*} = 1$ and preference distribution λ^d corresponds to $\psi^{\breve{\omega}*}(1, \lambda^d) = 1$. Proposition 4.2 applies and condition 4 of definition 5.6 is true.

Next, we turn to situation $\bar{\omega}$. Recall that $\gamma v(0) + \gamma\Delta k(0) < \theta_s^d\tilde{v}(0)$, $\gamma v(\frac{1}{2}) + \gamma\Delta k(\frac{1}{2}) > \theta_s^d\tilde{v}(\frac{1}{2})$, and $\Delta k', v', \tilde{v}' > 0$. Hence, $\gamma v(\phi^{\bar{\omega}}) + \gamma\Delta k(\phi^{\bar{\omega}})$ intersects $\theta_s^d\tilde{v}(\phi^{\bar{\omega}})$ from below at some $\phi^{\bar{\omega}} \in (0, \frac{1}{2})$. Let x be such an intersection. Thus,

- $\theta_s^d\tilde{v}(x) = \gamma v(x) + \gamma\Delta k(x) < \theta_s^d\tilde{v}(x) + \theta_s^d\tilde{h} + \theta_p^d \Rightarrow \theta_s^d = \frac{\gamma v(x) + \gamma\Delta k(x)}{\tilde{v}(x)}$, and
- $\theta_s^d\tilde{v}'(x) < \gamma v'(x) + \gamma\Delta k'(x) \Rightarrow \frac{\gamma v(x) + \gamma\Delta k(x)}{\tilde{v}(x)}\tilde{v}'(x) < \gamma v'(x) + \gamma\Delta k'(x)$.

By continuity of all involved functions, there is some $\epsilon > 0$ s.t. for all $y \in (x, x + \epsilon)$,

- $\theta_s^d\tilde{v}(y) < \gamma v(y) + \gamma\Delta k(y) < \theta_s^d\tilde{v}(y) + \theta_s^d\tilde{h} + \theta_p^d$,
- $\frac{\gamma v(y) + \gamma\Delta k(y)}{\tilde{v}(y)}\tilde{v}'(y) < \gamma v'(y) + \gamma\Delta k'(y)$, and

Consider any $\phi^{\bar{\omega}*} \in (x, x + \epsilon)$ and let the cost curve $-\Delta m^{\bar{\omega}}$ be s.t. $-\Delta m^{\bar{\omega}}(\phi^{\bar{\omega}*}) = \gamma v(\phi^{\bar{\omega}*}) + \gamma\Delta k(\phi^{\bar{\omega}*})$ and $-\Delta m^{\bar{\omega}'}(\phi^{\bar{\omega}*}) > \gamma v'(\phi^{\bar{\omega}*}) + \gamma\Delta k'(\phi^{\bar{\omega}*})$. Hence, we get that $\phi^{\bar{\omega}} < \frac{1}{2}$ satisfies

- $\theta_s^d\tilde{v}(\phi^{\bar{\omega}}) < -\Delta m^{\bar{\omega}}(\phi^{\bar{\omega}}) < \theta_s^d\tilde{v}(\phi^{\bar{\omega}}) + \theta_s^d\tilde{h} + \theta_p^d$,
- $\gamma v(\phi^{\bar{\omega}}) + \gamma\Delta k(\phi^{\bar{\omega}}) = -\Delta m^{\bar{\omega}}(\phi^{\bar{\omega}})$, and
- $\frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*})}{\tilde{v}(\phi_r^{\bar{\omega}*})} \times \frac{\mathrm{d}\tilde{v}(x)}{\mathrm{d}x}|_{x=\phi_r^{\omega*}} < \gamma(\frac{\mathrm{d}v(x)}{\mathrm{d}x}|_{x=\phi_r^{\omega*}} + \frac{\mathrm{d}\Delta k(x)}{\mathrm{d}x}|_{x=\phi_r^{\omega*}}) < -\frac{\mathrm{d}\Delta m^{\omega}(x)}{\mathrm{d}x}|_{x=\phi^{\omega}}$.

Thus, the cost curve $-\Delta m^{\bar{\omega}}$ is s.t. a cultural equilibrium of an imperfect social norm $\phi^{\bar{\omega}*} \in (0, 1)$ exists at λ^d. Moreover, note that $-\Delta m^{\breve{\omega}}(1) > \theta_s^d\tilde{v}(1) \Rightarrow \frac{-\Delta m^{\breve{\omega}}(1)}{\tilde{h}+\tilde{v}(1)}(\tilde{h} + \tilde{v}(1)) > \theta_s^d\tilde{v}(1)$ and $\frac{-\Delta m^{\breve{\omega}}(1)}{\tilde{h}+\tilde{v}(1)}(\tilde{h} + \tilde{v}(0)) > \theta_s^d\tilde{v}(0)$ imply that $\frac{-\Delta m^{\breve{\omega}}(1)}{\tilde{h}+\tilde{v}(1)}(\tilde{h} + \tilde{v}(\phi^{\bar{\omega}*})) > \theta_s^d\tilde{v}(\phi^{\bar{\omega}*})$.

Finally, consider any Ω s.t. the situations $\breve{\omega}$ and $\bar{\omega}$ are in it and have costs of contribution as above. Let ϕ_r^* be s.t. (1) $\phi_r^\omega = 0\ \forall \omega \in \Omega/\{\breve{\omega}, \bar{\omega}\}$, (2) $\phi_r^{\breve{\omega}*} = 1$, and (3) $\phi_r^{\bar{\omega}*} = \phi^{\bar{\omega}*} \in (0,1)$. The proposition is shown to be true by example. □

Proof of lemma 5.2:

Proof. Consider any ϕ_r^* of definition 5.6, $\Lambda_p(\phi_r^*)$ of definition 5.5, and any $\lambda \in \Lambda_P(\phi_r^*)$. By contradiction, assume that the lemma is not true. First, suppose condition 1 does not hold for some $\omega \in \Omega$ s.t. $\phi_r^{\omega*} \in (0,1)$. Note that $(\sigma_1^\omega, \sigma_0^\omega) = (1,0)$ for all $\sigma^\omega \in \Sigma^{\omega*}(\phi_r^{\omega*}, \lambda^d)$. Condition 3 of definition 5.6 and lemma 5.1 imply that for all $\theta \in \text{supp}(\lambda)$, $-\Delta m^\omega(\phi_r^{\omega*}) < \theta_s \tilde{v}(\phi_r^{\omega*}) + \theta_s \tilde{h} + \theta_p$. Since $\lambda \in \Lambda_p(\phi_r^*)$, $(\bar{\sigma}_1^\omega, \bar{\sigma}_0^\omega) = (1,0) \Rightarrow \theta_s \tilde{v}(\phi_r^{\omega*}) \leq -\Delta m^\omega(\phi_r^{\omega*})\ \forall \theta \in \text{supp}(\lambda)$. If $\omega \neq \text{argmin}\{\frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*})}{\tilde{v}(\phi_r^{\bar{\omega}*})} : \phi_r^{\bar{\omega}*} = \psi_r^{\bar{\omega}*}(\phi_r^{\bar{\omega}}, \lambda^d) \in (0,1)\}$, then $\theta_s \leq \min\{\frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*})}{\tilde{v}(\phi_r^{\bar{\omega}*})} : \phi_r^{\bar{\omega}*} = \psi^{\bar{\omega}*}(\phi_r^{\bar{\omega}*}, \lambda^d) \in (0,1)\} < \frac{-\Delta m^\omega(\phi_r^{\omega*})}{\tilde{v}(\phi_r^{\omega*})}$ for all $\theta \in \text{supp}(\lambda)$. Hence, $\theta_s \tilde{v}(\phi_r^{\omega*}) < -\Delta m^\omega(\phi_r^{\omega*})$. Alternatively, suppose $\omega = \text{argmin}\{\frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*})}{\tilde{v}(\phi_r^{\bar{\omega}*})} : \phi_r^{\bar{\omega}*} = \psi_r^{\bar{\omega}*}(\phi_r^{\bar{\omega}}, \lambda^d) \in (0,1)\}$. Hence condition 2 of definition 5.6 holds for this situation. From the above, we know that $\theta_s \tilde{v}(\phi_r^{\omega*}) \leq -\Delta m^\omega(\phi_r^{\omega*}) < \theta_s \tilde{v}(\phi_r^{\omega*}) + \theta_s \tilde{h} + \theta_p\ \forall \theta \in \text{supp}(\lambda)$. Assume by contradiction that there is some $\theta \in \text{supp}(\lambda)$ s.t. $-\Delta m^\omega(\phi_r^{\omega*}) = \theta_s \tilde{v}(\phi_r^{\omega*})$. Proposition 4.9 implies that $\phi_r^{\omega*}$ is not a cultural equilibrium at λ. Hence, $\lambda \notin \Lambda_p(\phi_r^*)$. We have reached a contradiction, implying that $\theta_s \tilde{v}(\phi_r^{\omega*}) < -\Delta m^\omega(\phi_r^{\omega*}) < \theta_s \tilde{v}(\phi_r^{\omega*}) + \theta_s \tilde{h} + \theta_p\ \forall \theta \in \text{supp}(\lambda)$.

Next, suppose that condition 2 of lemma 5.2 does not hold. Suppose condition 4a of definition 5.6 is true. Consequently, condition 2a of lemma 5.2 is true too. hence, 4a cannot be true. Thus, 4b must be true. Since $\theta_s < \min\{\frac{-\Delta m^{\bar{\omega}}(\phi_r^{\bar{\omega}*})}{\tilde{v}(\phi_r^{\bar{\omega}*})} : \phi_r^{\bar{\omega}*} = \psi_r^{\bar{\omega}*}(\phi_r^{\bar{\omega}}, \lambda^d) \in (0,1)\}$, condition 2b of lemma 5.2 is true. Hence, condition 2 of lemma 5.2 is true. □

Proof of lemma 5.4:

Proof. Consider any ϕ_r^* of definition 5.6, $\Lambda_r(\phi_r^*)$ of definition 5.5, $\lambda \in \Lambda_r(\phi_r^*)$ and $\hat{\lambda} \in [0,1]^{|\Theta|}$.

First, consider any $\omega \in \Omega$ s.t. $\phi_r^{\omega*} = 0$. $(\hat{\Phi}^{\omega*} = \{0\}$ and $\hat{\sigma}_n^\omega = \sigma_n^\omega\ \forall n \in \{0,1\}\backslash\{1 - \phi_r^{\omega*}\}, \hat{\sigma}^\omega \in \Sigma^\omega(\hat{\phi}^{\omega*}, \hat{\lambda}), \sigma^\omega \in \Sigma^\omega(\phi^{\omega*}, \lambda^d)) \Rightarrow \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}) = 0 = \psi^{\omega*}(0, \lambda)\ \forall \hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$. $\psi^{\omega*}(0, \hat{\lambda}) = 0 \Rightarrow \hat{\sigma}_{0,\theta}^\omega = 0 \Rightarrow B_\theta^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) = b^\omega(0,0,0,0) \Rightarrow B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) = B_{\hat{\lambda}}^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega)\ \forall \hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ and $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$.

Second, consider any $\omega \in \Omega$ s.t. $\phi_r^{\omega*} = 1$. $(\hat{\Phi}^{\omega*} = \{1\}$ and $\hat{\sigma}_n^\omega = \sigma_n^\omega\ \forall n \in \{0,1\}\backslash\{1 - \phi_r^{\omega*}\}, \hat{\sigma}^\omega \in \Sigma^\omega(\hat{\phi}^{\omega*}, \hat{\lambda}), \sigma^\omega \in \Sigma^\omega(\phi^{\omega*}, \lambda^d)) \Rightarrow \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}) = \psi^{\omega*}(1, \lambda) = 1\ \forall \hat{\phi}^\omega \in \hat{\Phi}^{\omega*} \Rightarrow$

$\hat{\sigma}^{\omega}_{1,\theta} = 1 \, \forall \theta \in \text{supp}(\hat{\lambda}), \hat{\sigma}^{\omega} \in \Sigma^{\omega *}(1, \hat{\lambda}) \Rightarrow B^{\omega}_{\lambda}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) = b^{\omega}(1,1,1,1) = B^{\omega}_{\hat{\lambda}}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) \, \forall \hat{\phi}^{\omega} \in \hat{\Phi}^{\omega *}$ and $\hat{\sigma}^{\omega} \in \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})$. Alternatively, suppose $\phi^{\omega *}_{r} = 1 > \psi^{\omega *}(1, \lambda)$. ($\hat{\Phi}^{\omega *} = \{1\}$ and $\hat{\sigma}^{\omega}_{n} = \sigma^{\omega}_{n} \; \forall n \in \{0,1\} \backslash \{1 - \phi^{\omega *}_{r}\}, \hat{\sigma}^{\omega} \in \Sigma^{\omega}(\hat{\phi}^{\omega *}, \hat{\lambda}), \sigma^{\omega} \in \Sigma^{\omega}(\phi^{\omega *}, \lambda^{d})) \Rightarrow \psi^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda}) = \psi^{\omega *}(1, \lambda^{d}) \in (0,1) \, \forall \hat{\phi}^{\omega} \in \hat{\Phi}^{\omega *} \Rightarrow \theta^{*}_{p} + \theta^{*}_{s} \tilde{h} + \theta^{d}_{s} \tilde{v}(1) = -\Delta m^{\omega}(\psi^{\omega *}(1, \hat{\lambda})) \Rightarrow b(1, 1, \psi^{\omega *}(1, \hat{\lambda}), 1) = b(0, 1, \psi^{\omega *}(1, \hat{\lambda}), 1)$ $\Rightarrow B^{\omega}_{\lambda}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) = B^{\omega}_{\hat{\lambda}}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) \, \forall \hat{\phi}^{\omega} \in \hat{\Phi}^{\omega *}$ and $\hat{\sigma}^{\omega} \in \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})$.

Lastly, consider any $\omega \in \Omega$ s.t. $\phi^{\omega *}_{r} = \psi^{\omega *}(1, \lambda) \in (0,1)$. ($\hat{\Phi}^{\omega *} = \{\phi^{\omega *}_{r}\}$ and $\hat{\sigma}^{\omega}_{n} = n \; \forall n \in \{0,1\}, \hat{\sigma}^{\omega} \in \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})) \Rightarrow (\hat{\sigma}^{\omega}_{1,\theta}, \hat{\sigma}^{\omega}_{0,\theta}) = (1,0) \, \forall \theta \in \text{supp}(\lambda) \cup \text{supp}(\hat{\lambda})$. Thus, $B^{\omega}_{\lambda}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) = \phi^{\omega *} b^{\omega}(1,1,\phi^{\omega *}, \phi^{\omega *}) + (1 - \phi^{\omega *}) b^{\omega}(0,0,\phi^{\omega *}, \phi^{\omega *}) = B^{\omega}_{\hat{\lambda}}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) \forall \hat{\phi}^{\omega} \in \hat{\Phi}^{\omega *}$ and $\hat{\sigma}^{\omega} \in \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})$.

The above shows that for any situation ω, $\lambda \in \Lambda_{r}(\phi^{*}_{r})$, and $\lambda \in [0,1]^{|\Theta|}$, the two stated conditions imply that $B^{\omega}_{\lambda}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) = B^{\omega}_{\hat{\lambda}}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) \, \forall \hat{\phi}^{\omega} \in \hat{\Phi}^{\omega *}$ and $\hat{\sigma}^{\omega} \in \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})$. It follows that the lemma is true. □

Proof of lemma 5.5:

Proof. Consider any ϕ^{*}_{r} of definition 5.6, $\Lambda_{r}(\phi^{*}_{r})$ of definition 5.5, $\lambda \in \Lambda_{r}(\phi^{*}_{r})$, and $\omega \in \Omega$ s.t. $\phi^{\omega *}_{r} = 1$. Lemma 5.3 implies that there is some neighborhood U of λ s.t. for all $\hat{\lambda} \in U^{\omega}$, $\phi^{\omega *}_{r} = 1$ is a cultural equilibrium. Hence, for all $\hat{\lambda} \in U$, $\hat{\Phi}^{\omega *} = \{1\}$. Throughout, consider any $\hat{\lambda} \in U$. The "or" condition of lemma 5.5 holds only if $\hat{\sigma}^{\omega}_{n} \neq = \sigma^{\omega}_{n}$ for some $n \in \{0,1\} \backslash \{1 - \phi^{\omega *}_{r}\} = \{1\}, \hat{\sigma}^{\omega} \in \Sigma^{\omega}(\hat{\phi}^{\omega *}, \hat{\lambda}), \sigma^{\omega} \in \Sigma^{\omega}(\phi^{\omega *}, \lambda^{d})$, which is true if and only if $\psi^{\omega *}(1, \hat{\lambda}) \neq \psi^{\omega *}(1, \lambda) = 1$.

First, suppose $\psi^{\omega *}(1, \hat{\lambda}) < \psi^{\omega *}(1, \lambda) = 1$. $\psi^{\omega *}(1, \lambda) = 1 \Rightarrow \psi^{\omega *}(1, \lambda^{d}) = 1 \Rightarrow \theta^{d}_{s} \tilde{v}(1) + \theta^{d}_{s} \tilde{h} + \theta^{d}_{p} > -\Delta m^{\omega}(1) \Rightarrow b^{\omega}(1,1,1,1) > b^{\omega}(0,1,1,1)$. $\psi^{\omega *}(1, \hat{\lambda}) < 1 \Rightarrow -\Delta m^{\omega}(\psi^{\omega *}(1, \hat{\lambda})) < -\Delta m^{\omega}(1) \Rightarrow b^{\omega}(1,1,\psi^{\omega *}(1, \hat{\lambda}), 1) - b^{\omega}(0,1,\psi^{\omega *}(1, \hat{\lambda}), 1) > b^{\omega}(1,1,1,1) - b^{\omega}(0,1,1,1) \geq 0$. Moreover, $\psi^{\omega *}(1, \lambda) = 1 \Rightarrow \theta_{p} + \theta_{s} \tilde{h} + \theta_{s} \tilde{v}(1) \geq -\Delta m^{\omega}(1) \forall \theta \in \text{supp}(\lambda) \Rightarrow \theta_{p} + \theta_{s} \tilde{h} + \theta_{s} \tilde{v}(1) > -\Delta m^{\omega}(\psi^{\omega *}(1, \hat{\lambda})) \forall \theta \in \text{supp}(\lambda) \Rightarrow \hat{\sigma}^{\omega}_{1,\theta} = 1 \forall \theta \in \text{supp}(\lambda), \hat{\sigma}^{\omega} \in \Sigma^{\omega *}(1, \hat{\lambda}) \Rightarrow B^{\omega}_{\lambda}(\hat{\sigma}^{\omega}, 1) = b^{\omega}(1,1,\psi^{\omega *}(1, \hat{\lambda}), 1) \, \forall \hat{\sigma}^{\omega} \in \Sigma^{\omega *}(1, \hat{\lambda})$. Therefore, $B^{\omega}_{\lambda}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) = b^{\omega}(1,1,\psi^{\omega *}(1, \hat{\lambda}), 1) > \psi^{\omega *}(1, \hat{\lambda}) b^{\omega}(1,1,\psi^{\omega *}(1, \hat{\lambda}), 1) + (1 - \psi^{\omega *}(1, \hat{\lambda})) b^{\omega}(0,1,\psi^{\omega *}(1, \hat{\lambda}), 1) = B^{\omega}_{\hat{\lambda}}(\hat{\sigma}^{\omega}, \hat{\phi}^{\omega}) \, \forall \hat{\phi}^{\omega} \in \hat{\Phi}^{\omega *}$ and $\hat{\sigma}^{\omega} \in \Sigma^{\omega *}(\hat{\phi}^{\omega}, \hat{\lambda})$.

Next, suppose $\phi^{\omega *} = 1 > \psi^{\omega *}(1, \lambda) > \psi^{\omega *}(1, \hat{\lambda}) \geq 0$. $\psi^{\omega *}(1, \lambda) \in (0,1) \Rightarrow \psi^{\omega *}(1, \lambda^{d}) \in (0,1) \Rightarrow \theta^{d}_{s} \tilde{v}(1) + \theta^{d}_{s} \tilde{h} + \theta^{d}_{p} = -\Delta m^{\omega}(\psi^{\omega *}(1, \lambda^{d})) \Rightarrow b^{\omega}(1,1,1,1) = b^{\omega}(0,1,1,1)$. $\psi^{\omega *}(1, \hat{\lambda}) < \psi^{\omega *}(1, \lambda) \Rightarrow -\Delta m^{\omega}(\psi^{\omega *}(1, \hat{\lambda})) < -\Delta m^{\omega}(\psi^{\omega *}(1, \lambda)) \Rightarrow b^{\omega}(1,1,\psi^{\omega *}(1, \hat{\lambda}), 1) - b^{\omega}(0,1,\psi^{\omega *}(1, \hat{\lambda}), 1)$

$> b^\omega(1,1,\psi^{\omega*}(1,\lambda),1) - b^\omega(0,1,\psi^{\omega*}(1,\lambda),1) = 0$. Moreover, $-\Delta m^\omega(\psi^{\omega*}(1,\hat{\lambda})) < -\Delta m^\omega(\psi^{\omega*}(1,\lambda))$ $\Rightarrow (\forall \theta \in \Theta, \theta_p + \theta_s\tilde{h} + \theta_s\tilde{v}(1) \geq -\Delta m^\omega(\psi^{\omega*}(1,\lambda)) \Rightarrow \theta_p + \theta_s\tilde{h} + \theta_s\tilde{v}(1) > -\Delta m^\omega(\psi^{\omega*}(1,\hat{\lambda}))) \Rightarrow$ $\hat{\sigma}_{1,\theta} \geq \sigma_{1,\theta}\, \forall \theta \in \text{supp}(\lambda) \Rightarrow \sum_{\theta\in\text{supp}(\lambda)} \lambda_\theta \hat{\sigma}_{1,\theta} \geq \sum_{\theta\in\text{supp}(\lambda)} \lambda_\theta \sigma_{1,\theta} = \psi^{\omega*}(1,\lambda) > \psi^{\omega*}(1,\hat{\lambda})\, \forall \sigma^\omega \in$ $\Sigma^{\omega*}(1,\lambda), \hat{\sigma}^\omega \in \Sigma^{\omega*}(1,\hat{\lambda})$. $(b^\omega(1,1,\psi^{\omega*}(1,\hat{\lambda}),1) > b^\omega(0,1,\psi^{\omega*}(1,\hat{\lambda}),1)$ and $\sum_{\theta\in\text{supp}(\lambda)} \lambda_\theta \hat{\sigma}_{1,\theta} >$ $\psi^{\omega*}(1,\hat{\lambda})\, \forall \sigma^\omega \in \Sigma^{\omega*}(1,\lambda), \hat{\sigma}^\omega \in \Sigma^{\omega*}(1,\hat{\lambda})) \Rightarrow B^\omega_\lambda(\hat{\sigma}^\omega, 1) = (\sum_{\theta\in\text{supp}(\lambda)} \lambda_\theta \hat{\sigma}_{1,\theta}) b^\omega(1,1,\psi^{\omega*}(1,\hat{\lambda}),1) +$ $(1 - \sum_{\theta\in\text{supp}(\lambda)} \lambda_\theta \hat{\sigma}_{1,\theta}) b^\omega(0,1,\psi^{\omega*}(1,\hat{\lambda}),1) > \psi^{\omega*}(1,\hat{\lambda}) \times b^\omega(1,1,\psi^{\omega*}(1,\hat{\lambda}),1) + (1 - \psi^{\omega*}(1,\hat{\lambda})) b^\omega(0,1,\psi^{\omega*}(1,\hat{\lambda}),$ $B^\omega_{\hat{\lambda}}(\hat{\sigma}^\omega, 1)\, \forall \hat{\sigma}^\omega \in \Sigma^{\omega*}(1,\hat{\lambda})$. Hence, $B^\omega_\lambda(\hat{\sigma}^\omega, \hat{\phi}^\omega) > B^\omega_{\hat{\lambda}}(\hat{\sigma}^\omega, \hat{\phi}^\omega)\, \forall \hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ and $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$. The case of $\phi^{\omega*} = 1 \geq \psi^{\omega*}(1,\hat{\lambda}) > \psi^{\omega*}(1,\lambda) > 0$ works analogously to the above. Therefore, we refrain from writing it out.

We have shown that the lemma holds for all different cases of $\psi^{\omega*}(1,\lambda)$ and $\psi^{\omega*}(1,\hat{\lambda})$. Hence, it must be true. □

Lemma B.15. *Consider any ϕ^*_r of definition 5.6, $\Lambda_r(\phi^*_r)$ of definition 5.5, and any $\omega \in \Omega$ s.t. $\phi^{\omega*}_r \in (0,1)$. Let $\bar{\delta} > 0$ be s.t. proposition B.1 applies. For all $\lambda \in \Lambda_r(\phi^*_r)$, there is some $\delta \in (0, \bar{\delta})$ s.t. for all $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta} |\lambda_\theta - x_\theta| < \delta\}$, there is some cultural equilibrium $\hat{\Phi}^{\omega*}$ at $\hat{\lambda}$ s.t. for all $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ and $\sigma \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$, $\hat{\phi}^\omega \neq \phi^{\omega*} \Rightarrow (\hat{\sigma}^\omega_1, \hat{\sigma}^\omega_0) \neq (1,0)$.*

Proof. Consider any case as described above. Let δ be so small that for all $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$, $\hat{\phi}^\omega \neq \phi^{\omega*}_r \Rightarrow (\sigma^\omega_1, \sigma^\omega_0) = (\bar{\sigma}^\omega_1, \bar{\sigma}^\omega_0)$ for all $\sigma^\omega, \bar{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$. To see that such a δ exists, let $\epsilon > 0$ be s.t. for all $x \in (\phi^{\omega*}_r - \epsilon, \phi^{\omega*}_r) \cup (\phi^{\omega*}_r, \phi^{\omega*}_r + \epsilon)$, $\hat{n}, \check{n} \in \{0,1\}$, and $\hat{\theta}, \check{\theta} \in \Theta$: $\hat{\theta}_s\tilde{v}(\phi^{\omega*}_r) + \hat{n}(\hat{\theta}_s\tilde{h} + \hat{\theta}_p) > \check{\theta}_s\tilde{v}(\phi^{\omega*}_r) + \check{n}(\check{\theta}_s\tilde{h} + \check{\theta}_p) \Rightarrow \hat{\theta}_s\tilde{v}(x) + \hat{n}(\hat{\theta}_s\tilde{h} + \hat{\theta}_p) > \check{\theta}_s\tilde{v}(x) + \check{n}(\check{\theta}_s\tilde{h} + \check{\theta}_p)$. Moreover, note that for all $\hat{\theta}$ and $\check{\theta}$, $\hat{\theta}_s\tilde{v}(\phi^{\omega*}_r) + \hat{\theta}_s\tilde{h} + \hat{\theta}_p = \check{\theta}_s\tilde{v}(\phi^{\omega*}_r) \Rightarrow \hat{\theta}_s\tilde{v}(x) + \hat{\theta}_s\tilde{h} + \hat{\theta}_p \neq \check{\theta}_s\tilde{v}(x)\, \forall x \in (\phi^{\omega*}_r - \epsilon, \phi^{\omega*}_r) \cup (\phi^{\omega*}_r, \phi^{\omega*}_r + \epsilon)$. Hence, $\hat{\theta}_p + \hat{\theta}_s\tilde{h} + \hat{\theta}_s\tilde{v}(\phi^\omega) \neq \tilde{\theta}_s\tilde{v}(\phi^\omega)\ \forall \hat{\theta}, \tilde{\theta} \in \text{supp}(\lambda)$. Lemma B.7 then implies that $(\sigma^\omega_1, \sigma^\omega_0) = (\bar{\sigma}^\omega_1, \bar{\sigma}^\omega_0)$ for all $\sigma^\omega, \bar{\sigma}^\omega \in \Sigma^{\omega*}(x, \hat{\lambda})$ and $x \in\in (\phi^{\omega*}_r - \epsilon, \phi^{\omega*}_r) \cup (\phi^{\omega*}_r, \phi^{\omega*}_r + \epsilon)$. Finally, let δ be sufficiently small s.t. for all $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$, $\hat{\phi}^\omega \in (\phi^{\omega*}_r - \epsilon, \phi^{\omega*}_r) \cup (\phi^{\omega*}_r, \phi^{\omega*}_r + \epsilon)$. Such a δ exists by proposition 4.8.

Assume by contraction that $\exists \hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ s.t. $\hat{\phi}^\omega \neq \phi^{\omega*}_r$ and $(\hat{\sigma}^\omega_1, \hat{\sigma}^\omega_0) = (1,0)$ for some $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$. Note that since $\Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$ is a singleton, the second condition can be rewritten as $(\hat{\sigma}^\omega_1, \hat{\sigma}^\omega_0) = (1,0)$ for all $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$.

Consider some $\eta > 0$ s.t. $\forall x \in (\phi^{\omega*} - \eta, \phi^{\omega*}) \cup (\phi^{\omega*},)\phi^{\omega*} + \eta)$ (1) $\frac{\dot{x}}{\phi^{\omega*} - x} > 0$ at λ and (2) $(\check{\sigma}^\omega_1, \check{\sigma}^\omega_0) = (1,0) \forall \check{\sigma}^\omega \in \Sigma^{\omega*}(x, \lambda)$. Such an η exists by the same reasoning as in

the proof of proposition B.1. $(\frac{\dot{x}}{\phi^{\omega*}-x} > 0$ at λ and $(\check{\sigma}_1^\omega, \check{\sigma}_0^\omega) = (1,0) \forall \check{\sigma}^\omega \in \Sigma^{\omega*}(x,\lambda)) \Rightarrow \frac{\gamma v(x) + \Delta m^\omega(x,\lambda) + \gamma \Delta k(x)}{\phi^{\omega*}-x} > 0$. $(\hat{\phi}^\omega \neq \phi^{\omega*}$ and $(\hat{\sigma}_1^\omega, \hat{\sigma}_0^\omega) = (1,0)) \Rightarrow \frac{\gamma v(\hat{\phi}^\omega + \Delta m^\omega(\hat{\phi}^\omega,\lambda) + \gamma \Delta k(\hat{\phi}}{\phi^{\omega*}-\hat{\phi}} > 0 \Rightarrow \dot{\hat{\phi}}^\omega \neq 0$. Hence, $\hat{\phi}^\omega \notin \hat{\Phi}^{\omega*}$, since it is not a rest point. We have reached a contradiction. Thus, lemma B.15 is true. □

Proof of lemma 5.6:

Proof. Consider any ϕ_r^* of definition 5.6, $\Lambda_r(\phi_r^*)$ of definition 5.5, $\lambda \in \Lambda_r(\phi_r^*)$, and $\omega \in \Omega$ s.t. $\phi_r^{\omega*} \in (0,1)$. Let the neighborhood U be s.t. proposition B.1 and lemma B.15 apply. Consider any $\hat{\lambda} \in U$ s.t. (1) $\phi_r^{\omega*} \notin \hat{\Phi}^{\omega*}$ or (2) $\exists n \in \{0,1\}$ s.t. $\hat{\sigma}_n^\omega \neq n \, \forall \hat{\sigma}^\omega \in \Sigma^{\omega*}(\phi_r^{\omega*}, \hat{\lambda})$. Lemma B.15 implies that for all $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$, $(\hat{\phi}^\omega \neq \phi_r^{\omega*} \Rightarrow \exists n \in \{0,1\}$ s.t. $\hat{\sigma}_n^\omega \neq n \, \forall \sigma^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}))$. Hence, $\exists n \in \{0,1\}$ s.t. $\hat{\sigma}_n^\omega \neq n \, \forall \hat{\phi}^\omega \in \hat{\Phi}^{\omega*}, \hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$. Below, consider any $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ and $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$. Since proposition B.1 applies:

- $\theta_s \tilde{v}(\hat{\phi}^\omega) < -\Delta m^\omega(\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})) < \theta_p + \theta_s \tilde{h} + \theta_s \tilde{v}(\hat{\phi}^\omega) \; \forall \theta \in \text{supp}(\lambda)$ and
- $b^\omega(n, n, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) > b^\omega(1-n, n, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) \; \forall n \in \{0,1\}$.

$\theta_s \tilde{v}(\hat{\phi}^\omega) < -\Delta m^\omega(\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})) < \theta_p + \theta_s \tilde{h} + \theta_s \tilde{v}(\hat{\phi}^\omega) \, \forall \theta \in \text{supp}(\lambda) \Rightarrow (\hat{\sigma}_{1,\theta}^\omega, \hat{\sigma}_{0,\theta}^\omega) = (1,0) \, \forall \theta \in \text{supp}(\lambda), \hat{\sigma}^\omega \Rightarrow B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) = \hat{\phi}^\omega b^\omega(1,1,\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) + (1-\hat{\phi}^\omega) b^\omega(0,0,\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega)$. $(b^\omega(n,n,\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) > b^\omega(1-n,n,\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) \; \forall n \in \{0,1\}$ and $(\hat{\sigma}_1^\omega, \hat{\sigma}_0^\omega) \neq (1,0)) \Rightarrow B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) > \hat{\phi}^\omega \hat{\sigma}_1^\omega b^\omega(1,1,\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) + \hat{\phi}^\omega(1-\hat{\sigma}_1^\omega) b^\omega(0,1,\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) + (1-\hat{\phi}^\omega)\hat{\sigma}_0^\omega \times b^\omega(1,0,\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) + (1-\hat{\phi}^\omega)(1-\hat{\sigma}_0^\omega) b^\omega(0,0,\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) = B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega)$. □

Lemma B.16. *Consider any ϕ_r^* of definition 5.6 and $\Lambda_p(\phi_r^*)$ of definition 5.5. For each preference distribution $\hat{\lambda} \in [0,1]^{|\Theta|}$, let $\hat{\Phi}^{\omega*}$ be the cultural equilibrium at $\hat{\lambda}$ s.t. ϕ_r^* is in it's basin of attraction.*

For all $\lambda \in \Lambda_p(\phi_r^*)$, there is some neighborhood U of λ s.t. for all $\hat{\lambda} \in U$ and $\omega \in \{x \in \Omega : \phi_r^{x*} \in (0,1)\}$: $B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) \geq B_{\hat{\lambda}}^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega)$ for all $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ and $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$.

Proof. Consider any ϕ_r^* of definition 5.6, $\Lambda_r(\phi_r^*)$ of definition 5.5, $\lambda \in \Lambda_r(\phi_r^*)$, and $\omega \in \Omega$ s.t. $\phi_r^{\omega*} \in (0,1)$. Let the neighborhood U be s.t. proposition B.1 applies. Consider any $\hat{\lambda} \in U$, $\hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$, and $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$. Since proposition B.1 applies:

- $\theta_s\tilde{v}(\hat{\phi}^\omega) < -\Delta m^\omega(\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})) < \theta_p + \theta_s\tilde{h} + \theta_s\tilde{v}(\hat{\phi}^\omega)\ \forall\theta \in \text{supp}(\lambda)$ and
- $b^\omega(n, n, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) > b^\omega(1-n, n, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega)\ \forall n \in \{0, 1\}$.

$\theta_s\tilde{v}(\hat{\phi}^\omega) < -\Delta m^\omega(\psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})) < \theta_p + \theta_s\tilde{h} + \theta_s\tilde{v}(\hat{\phi}^\omega)\,\forall\theta \in \text{supp}(\lambda) \Rightarrow (\hat{\sigma}^\omega_{1,\theta}, \hat{\sigma}^\omega_{0,\theta}) = (1, 0)\,\forall\theta \in \text{supp}(\lambda), \hat{\sigma}^\omega \Rightarrow B^\omega_\lambda(\hat{\sigma}^\omega, \hat{\phi}^\omega) = \hat{\phi}^\omega b^\omega(1, 1, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) + (1-\hat{\phi}^\omega)b^\omega(0, 0, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega)$. $(b^\omega(n, n, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) > b^\omega(1-n, n, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega)\ \forall n \in \{0, 1\} \Rightarrow B^\omega_\lambda(\hat{\sigma}^\omega, \hat{\phi}^\omega) \geq \hat{\phi}^\omega x \times b^\omega(1, 1, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) + \hat{\phi}^\omega(1-x)b^\omega(0, 1, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) + (1-\hat{\phi}^\omega)yb^\omega(1, 0, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega) + (1-\hat{\phi}^\omega)(1-y)b^\omega(0, 0, \psi^{\omega*}(\hat{\phi}^\omega, \hat{\lambda}), \hat{\phi}^\omega)\,\forall x, y \in [0, 1] \geq B^\omega_{\tilde{\lambda}}(\hat{\sigma}^\omega, \hat{\phi}^\omega)$. □

Lemma B.17. *Consider any ϕ^*_r of definition 5.6 and $\Lambda_p(\phi^*_r)$ of definition 5.5. For each preference distribution $\hat{\lambda} \in [0, 1]^{|\Theta|}$, let $\hat{\Phi}^{\omega*}$ be the cultural equilibrium at $\hat{\lambda}$ s.t. ϕ^*_r is in it's basin of attraction.*

For all $\lambda \in \Lambda_p(\phi^*_r)$, there is some neighborhood U of λ s.t. for all $\hat{\lambda} \in U$ and $\omega \in \{x \in \Omega : \phi^{x*}_r \in (0, 1)\}$: If

1. *$\hat{\Phi}^{\omega*} \neq \{\phi^{\omega*}_r\}$ but $\phi^{\omega*}_r \in \hat{\Phi}^{\omega*}$ or*
2. *$(\hat{\sigma}^\omega_1, \hat{\sigma}^\omega_0) \neq (\sigma^\omega_1, \sigma^\omega_0)$ for some but not all $\hat{\sigma}^\omega \in \Sigma^\omega(\phi^*_r, \hat{\lambda})$ and $\sigma^\omega \in \Sigma^\omega(\phi^*_r, \lambda^d)$,*

then $\exists\bar{\omega} \in \Omega$ for which lemma 5.5 or lemma 5.6 applies.

Proof. Consider any ϕ^*_r of definition 5.6, $\Lambda_r(\phi^*_r)$ of definition 5.5, $\lambda \in \Lambda_r(\phi^*_r)$, and $\omega \in \Omega$ s.t. $\phi^{\omega*}_r \in (0, 1)$. Let U be so small that for all $\omega \in \Omega$ and $\hat{\lambda} \in U$, proposition B.1 applies if $\phi^{\omega*}_r \in (0, 1)$ and proposition 4.6 applies if $\phi^{\omega*}_r = 1$. Since $\phi^{\omega*}_r$ is a cultural equilibrium of proposition 4.3 at preference distribution λ^d, $\tilde{v}(\phi^{\omega*}_r) + \Delta m^\omega(\phi^{\omega*}_r) + \Delta k(\phi^{\omega*}_r) = 0$.

Suppose $\phi^{\omega*}_r \in \hat{\Phi}^{\omega*} \Rightarrow \dot{\phi}^{\omega*}_r$ at $\hat{\lambda} \Rightarrow (\hat{\sigma}^\omega_1 - \hat{\sigma}^\omega_0))\tilde{v}(\phi^{\omega*}_r) + \Delta m^\omega(\phi^{\omega*}_r)) + \Delta k(\phi^{\omega*}_r) = 0$ for some $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\phi^{\omega*}_r, \hat{\lambda})$. Since $\tilde{v}(\phi^{\omega*}_r) + \Delta m^\omega(\phi^{\omega*}_r) + \Delta k(\phi^{\omega*}_r) = 0$, it follows that $(\hat{\sigma}^\omega_1, \hat{\sigma}^\omega_0) = (1, 0)$ for some $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\phi^{\omega*}_r, \hat{\lambda})$. Given this equilibrium behavior, $\phi^{\omega*}_r$ is not a cultural equilibrium if and only if $\exists\theta \in \text{supp}(\hat{\lambda}), n \in \{0, 1\}$ s.t. $\theta_s\tilde{v}(\phi^{\omega*}_r) + n(\theta_s\tilde{h} + \theta_p) = -\Delta m^\omega(\phi^{\omega*}_r)$. Otherwise, $\phi^{\omega*}_r$ would be a cultural equilibrium at $\hat{\lambda}$ by proposition 4.7.

Similarly, $(\hat{\sigma}^\omega_1, \hat{\sigma}^\omega_0) \neq (\sigma^\omega_1, \sigma^\omega_0)$ for some but not all $\hat{\sigma}^\omega \in \Sigma^\omega(\phi^*_r, \hat{\lambda}), \sigma^\omega \in \Sigma^\omega(\phi^*_r, \lambda^d)$ also implies that there is $\theta \in \text{supp}(\hat{\lambda})$ and $n \in \{0, 1\}$ s.t. $\theta_s\tilde{v}(\phi^{\omega*}_r) + n(\theta_s\tilde{h} + \theta_p) = -\Delta m^\omega(\phi^{\omega*}_r)$

(follows from lemma B.7). Hence, if condition 1 or 2 holds, then $\exists \theta \in \text{supp}(\hat{\lambda}), n \in \{0,1\}$ s.t. $\theta_s \tilde{v}(\phi_r^{\omega*}) + n(\theta_s \tilde{h} + \theta_p) = -\Delta m^\omega(\phi_r^{\omega*})$.

First, suppose that for some $\theta \in \text{supp}(\hat{\lambda})$, $\theta_s \tilde{v}(\phi_r^{\omega*}) + \theta_s \tilde{h} + \theta_p = -\Delta m^\omega(\phi_r^{\omega*})$. This is only possible if $\theta_s \tilde{v}(1) + \theta_s \tilde{h} + \theta_p < \max_{\omega \in \Omega}\{-\Delta m^\omega(1) : \phi_r^{\omega*} = \psi^{\omega*}(1, \lambda^d) = 1\}$. Hence, for $\bar{\omega} := \text{argmax}_{\omega \in \Omega}\{-\Delta m^\omega(1) : \phi_r^{\omega*} = \psi^{\omega*}(1, \lambda^d) = 1\}$, $\psi^{\bar{\omega}*}(1, \hat{\lambda}) < 1 = \psi^{\bar{\omega}*}(1, \lambda)$. Lemma 5.5 applies.

Second, suppose that for some $\theta \in \text{supp}(\hat{\lambda})$, $\theta_s \tilde{v}(\phi_r^{\omega*}) = -\Delta m^\omega(\phi_r^{\omega*})$. Let $\breve{\omega} := \text{argmin}_{\omega \in \Omega}\{\frac{-\Delta m^\omega(\phi_r^{\omega*})}{\tilde{v}(\phi_r^{\omega*})} : \phi_r^{\omega*} \in (0,1)\}$. Note that $\omega \neq \breve{\omega}$, since $\theta_s \tilde{v}(\phi_r^{\omega*}) = -\Delta m^\omega(\phi_r^{\omega*})$ for some $\theta \in \text{supp}(\hat{\lambda})$, condition 2 of definition 5.6, and proposition 4.9 imply that $\phi_r^{\omega*} \notin \hat{\Phi}^{\omega*}$. $\omega \neq \breve{\omega}$ and $\theta_s \tilde{v}(\phi_r^{\omega*}) = -\Delta m^\omega(\phi_r^{\omega*})$ for some $\theta \in \text{supp}(\hat{\lambda})$ imply that $\theta_s \tilde{v}(\phi_r^{\breve{\omega}*}) < -\Delta m^{\breve{\omega}}(\phi_r^{\omega*})$. Hence, $(\hat{\sigma}_1^{\breve{\omega}}, \hat{\sigma}_0^{\breve{\omega}}) \neq (1,0) \, \forall \hat{\sigma}^\omega \in \Sigma^{\omega*}(\phi_r^{\breve{\omega}*}, \hat{\lambda})$. Thus, the stated conditions of lemma 5.6 hold for situation $\breve{\omega}$. □

Lemma B.18. *Consider any ϕ_r^* and $\Lambda_p(\phi^*)$ satisfying definitions 5.6 and 5.5 respectively. For all $\lambda \in \Lambda_p(\phi^*)$ there is some neighborhood U of λ s.t. $\hat{\lambda} \in U / \Lambda_p(\phi_r^*) \Rightarrow B_\lambda(\hat{\sigma}, \hat{\phi}) > B_{\hat{\lambda}}(\hat{\sigma}, \hat{\phi}) \, \forall \hat{\phi}^\omega \in \prod_{\omega \in \Omega} \hat{\Phi}^{\omega*}$ and $\hat{\sigma}^\omega \in \prod_{\omega \in \Omega} \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$.*

Proof. Consider any ϕ_r^* of definition 5.6, $\Lambda_r(\phi_r^*)$ of definition 5.5, $\lambda \in \Lambda_r(\phi_r^*)$. Let the neighborhood U of λ be s.t. (1) for all $\omega \in \{x \in \Omega, \phi_r^{x*} = 1\}$, either lemma 5.4 or lemma 5.5 applies and (2) for all $\omega \in \{x \in \Omega, \phi_r^{x*} \in (0,1)\}$ either lemma 5.4, lemma 5.6, or lemma B.16 applies. It follows that for all $\omega \in \Omega$, $B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) \geq B_{\hat{\lambda}}^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) \, \forall \hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ and $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$.

Since $\hat{\lambda} \notin \Lambda_p(\phi_r^*)$, there is some $\omega \in \Omega$ s.t. (1) $\hat{\Phi}^{\omega*} = \{\phi_r^{\omega*}\}$ or (2) $(\hat{\sigma}_1^\omega, \hat{\sigma}_0^\omega) \neq (\sigma_1^\omega, \sigma_0^\omega)$ for some $n \in \{0,1\} \backslash \{1 - \phi_r^{\omega*}\}, \hat{\sigma}^\omega \in \Sigma^\omega(\hat{\phi}^{\omega*}, \hat{\lambda}), \sigma^\omega \in \Sigma^\omega(\phi^{\omega*}, \lambda^d)$. First, suppose $\phi_r^{\omega*} = 1$. In that case, lemma 5.5 implies that $B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) > B_{\hat{\lambda}}^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) \, \forall \hat{\phi}^\omega \in \hat{\Phi}^{\omega*}$ and $\hat{\sigma}^\omega \in \Sigma^{\omega*}(\hat{\phi}^\omega, \hat{\lambda})$. Next, suppose $\phi_r^{\omega*} \in (0,1)$. In that case, lemma B.17 implies that there is some $\bar{\omega}$ for which lemma 5.5 or lemma 5.6 applies. Hence, $B_\lambda^{\bar{\omega}}(\hat{\sigma}^{\bar{\omega}}, \hat{\phi}^{\bar{\omega}}) > B_{\hat{\lambda}}^{\bar{\omega}}(\hat{\sigma}^{\bar{\omega}}, \hat{\phi}^{\bar{\omega}}) \, \forall \hat{\phi}^{\bar{\omega}} \in \hat{\Phi}^{\bar{\omega}*}$ and $\hat{\sigma}^{\bar{\omega}} \in \Sigma^{\bar{\omega}*}(\hat{\phi}^{\bar{\omega}}, \hat{\lambda})$. We can summarize the above as follows:

- for all $\omega \in \Omega$, $B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) \geq B_{\hat{\lambda}}^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) \, \forall \hat{\phi}^{\bar{\omega}} \in \hat{\Phi}^{\bar{\omega}*}$ and $\hat{\sigma}^{\bar{\omega}} \in \Sigma^{\bar{\omega}*}(\hat{\phi}^{\bar{\omega}}, \hat{\lambda})$ and
- for some $\omega \in \Omega$, $B_\lambda^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) > B_{\hat{\lambda}}^\omega(\hat{\sigma}^\omega, \hat{\phi}^\omega) \, \hat{\phi}^{\bar{\omega}} \in \hat{\Phi}^{\bar{\omega}*}$ and $\hat{\sigma}^{\bar{\omega}} \in \Sigma^{\bar{\omega}*}(\hat{\phi}^{\bar{\omega}}, \hat{\lambda})$,

implying that $B_\lambda(\hat{\sigma},\hat{\phi}) > B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi})\,\forall\hat{\phi}^\omega \in \prod_{\omega\in\Omega}\hat{\Phi}^{\omega*}$ and $\hat{\sigma}^\omega \in \prod_{\omega\in\Omega}\Sigma^{\omega*}(\hat{\phi}^\omega,\hat{\lambda})$. Hence, proposition B.18 is true. □

Proof of Proposition 5.1:

Proof. Consider any ϕ_r^* and $\Lambda_p(\phi_r^*)$ satisfying definitions 5.6 and 5.5 respectively. Moreover, consider any $\lambda \in \Lambda_p(\phi_r^*)$. Let U be s.t. lemma B.18 applies. Consider any $\hat{\lambda} \in U$. At preference distribution $\hat{\lambda}$, norms and behavior reach equilibria before further changes in the preference distribution occur. Moreover, lemma B.18 implies that at any social norms $\hat{\phi} \in \prod_{\omega\in\Omega}\hat{\Phi}^{\omega*}$ and behavior $\hat{\sigma} \in \prod_{\omega\in\Omega}\hat{\Sigma}^{\omega*}(\hat{\phi}^\omega,\hat{\lambda})$ in equilibrium, $B_\lambda(\hat{\sigma},\hat{\phi}) > B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi})$. Following Weibull (1997), this condition ensures that on the dynamic system of all $\theta \in \text{supp}(\hat{\lambda})$, preferences evolve towards some $\check{\lambda} \in \Lambda_p(\phi^*)$. Throughout, the course of preference evolution, the perfect social norm remains a cultural equilibrium at any ω s.t. $\phi_r^{\omega*} = 1$ (see proposition 4.6). Moreover, at any ω s.t. $\phi_r^{\omega*} \in (0,1)$, the cultural equilibrium remains so close to $\phi_r^{\omega*}$ that at any $\lambda \in \lambda^d$ the social norm returns to $\phi_r^{\omega*}$ (see proposition B.1). Consequently, once preferences return to $\Lambda_p(\phi^*)$, the social norms return to ϕ_r^*. At any preference distribution $\tilde{\lambda} \in \Lambda_p(\phi^*)$ and social norms ϕ_r^*, equilibrium behavior reaches $\Sigma^{\omega*}(\phi^{\omega*},\tilde{\lambda})$ in each situation ω, since it is the unique behavioral equilibrium (see proposition 3.1). Hence, proposition 5.1 is true. □