Games and Economic Behavior 147 (2024) 242–267

Contents lists available at ScienceDirect

Games and Economic Behavior

journal homepage: www.elsevier.com/locate/geb

Cooperation, norms, and gene-culture coevolution ☆,☆☆

Fabian Mankat

University of Kassel, Nora-Platiel-Straße 4, Kassel, 34109, Germany

ARTICLE INFO

JEL classification:
C73
D02
D91
Z13

Keywords:
Evolutionary game theory
Cooperation
Norms

ABSTRACT

This paper studies how a continuum of individuals interacting in a binary public goods game can secure cooperation through transmitting and enforcing norms. The evolutionary model consists of three distinct dimensions: behavior, norms, and approval preferences. In line with the *indirect approach* proposed by Güth and Yaari (1992), behavior results from utility maximization, while norms and approval preferences evolve over time. The underlying evolutionary processes differ concerning speed and nature. Whereas norms evolve at the cultural level through peer interactions and socialization, approval preferences are (at least partly) biologically inherited and transmitted from parents to their offspring. We find that if cultural and biological reproductive fitnesses are derived from material and social factors, then an interplay of social disapproval mechanisms gives rise to stable equilibria in which positive cooperation levels persist. Moreover, we find stable equilibria characterized by heterogeneous behavior and moral attitudes across individuals.

1. Introduction

Human societies uphold cooperation among non-related individuals, even if such behavior is relatively costly to the individuals themselves. The ability to bridge the divergence of self-interest and cooperation is often accredited to the existence and transmission of informal institutions such as social norms (see, 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 (Voss, 2001; Fehr and Fischbacher, 2004). 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.[1] However, it raises new questions regarding their underlying evolutionary foundations.

This paper investigates how a continuum of individuals recurrently interacting in a binary public goods game can secure cooperation through the transmission of a cooperation-prescribing norm and norm-sensitive preferences. To this end, we identify and study dynamically stable equilibria. The main contribution to the existing literature is two-fold. First, to the best of my knowledge, the paper presents the first model that describes the co-evolution of personal norms, social norms, and approval preferences. Thereby, it

☆ The author declares no competing interests.

☆☆ *Abbreviations:* NE, Nash equilibrium; CE, cultural equilibrium; BE, biological equilibrium.

E-mail address: fabian.mankat@uni-kassel.de.

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

https://doi.org/10.1016/j.geb.2024.07.006
Received 24 April 2023

Available online 10 August 2024

endogenizes the formation of norms and the mechanism that enforces them. Second, we incorporate a variety of different social disapproval mechanisms and thereby highlight their potential role. We find that an interplay of social disapproval mechanisms enables the existence of stable equilibria with positive cooperation levels as well as heterogeneous behavior, personal norms, and approval preferences across individuals.

The model consists of three distinct dimensions: behavior, norms, and approval preferences. Following the *indirect approach* proposed by Güth and Yaari (1992), individuals act rationally and maximize utility. Norms and approval preferences are subject to evolution. Norms evolve at the cultural level through peer interactions and socialization institutions (Henrich and Gil-White, 2001). Approval preferences are (at least partly) transmitted through biological reproduction (Chudek and 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 (Turke, 1989; Buss and Schmitt, 1993). Thus, their personal norms and approval preferences spread in society. The distribution of personal norms specifies what the individuals generally regard as appropriate, thus defining the social norm (Cooter, 1998; Carbonara et al., 2008). Social disapproval arises from three sources: social norm violation by acting against what society generally considers appropriate, personal norm non-conformity by holding moral views that conflict with those of others, and hypocrisy by engaging in behavior that conflicts with one's own personal norm.

The paper contributes to the evolutionary literature on norms employing the *indirect evolutionary approach* proposed by Güth and Yaari (1992). The underlying idea is that utility governs behavior, determining the reproductive fitness of cultural and biological traits that, in turn, shape the utility function.[2]

Mengel (2008) uses this approach to analyze the cultural transmission of cooperation norms in non-integrated societies. Individuals are recurrently matched to interact in the prisoners' dilemma. Any individual having internalized the cooperation norm experiences internal sanctions when she defects. Incomplete integration is modeled through a biased matching structure that favors interaction between alike individuals (in terms of norm internalization). Mengel (2008) finds that cooperation norms of intermediate strength can survive at 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, 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.

Fershtman and Weiss (1998) study another evolutionary model of norm-sensitive preferences. Individuals are recurrently matched into pairs and interact in a continuous contribution game. Fershtman and Weiss (1998) focus on social rewards as a behavioral co-determinant, where social rewards for complying with the social norm positively depend on the average contribution level in society. Their results indicate that if preferences are observable, individuals condition their behavior based on the preferences of the player they are facing, which enables preferences for social rewards to be evolutionary stable.

The model of this paper complements the above analyses by studying situations where the material payoff does not depend on the behavior of in-group peers but rather on that of society as a whole. In such situations, assortative matching and observable preferences provide 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 disutility of sanctions for social norm violation. Similar to Fershtman and Weiss (1998), sanctions positively depend on society's overall contribution level. This setup gives rise to multiple behavioral equilibria. Society reaches each behavioral equilibrium with some positive probability for every interaction in the public goods game. Similar to the model of this paper, 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 their behavior to the given environment.

The remainder of this paper is as follows. Section 2 presents the static theoretical framework. Section 3 discusses equilibrium behavior for exogenous norms and preferences. Sections 4 and 5 study the evolution of norms and preferences, respectively. Section 6 discusses the results, whilst Section 7 concludes and provides an outlook for future research.

2. Theoretical framework

We study a continuum of individuals $i \in \mathcal{I} = [0,1]$, who recurrently interact in a binary public goods game. Each individual i executes action $a^i \in \mathcal{A} = \{0,1\}$, which corresponds to either contributing to the public good ($a^i = 1$, 'cooperate') or not ($a^i = 0$, 'defect'). The share of individuals that contribute is ψ. At times, we refer to ψ as the cooperation level or share.

[2] Many other strands of literature look at norms in an evolutionary context. Young (1993, 1996, 2015), Sethi and Somanathan (1996), Binmore and Samuelson (1994), Nyborg and Rege (2003b), Rege (2004), Azar (2004), and Lindbeck et al. (1999) focus on the evolution of behavior to rationalize norm-compliance. Bisin and Verdier (1998); Bisin et al. (2004); Bisin and Verdier (2001), Tabellini (2008), and Bezin (2019) 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 the persuasion of peers. Boyd and Richerson (1990), Bowles and Gintis (1998), Mitteldorf and Wilson (2000), Henrich (2004), and Boyd and Richerson (2005) 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 (Bester and Güth, 1998; Guttman, 2003, 2013; Poulsen and Poulsen, 2006; Müller and von Wangenheim, 2019).

An individual i's approval preferences are indicated by her preference type $\theta^i = (\theta^i_s, \theta^i_p) \in \Theta \subset \mathbb{R}^2_{\geq 0}$, 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 a distribution λ indicates the existing preference types in this population, $\text{supp}(\lambda) := \{\theta \in \Theta : \lambda_\theta > 0\}$. We require that $\sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta = 1$.

We write an individual i's personal norm as $n^i \in \{0,1\}$. Individual i holds the *cooperation norm*, $n^i = 1$, if she considers cooperation the only morally right thing to do. Suppose individual i does not hold the cooperation norm, $n^i = 0$. In that case, she considers all possible actions appropriate.[4] We sometimes refer to the sub-population of individuals who hold the cooperation norm as *norm holders*. Analogously, an individual is a *norm non-holder* if she does not hold the cooperation norm. Individuals communicate their personal norms to peers. We assume 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]

The social norm captures societies' shared understanding of appropriate behavior. 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, the distribution of personal norms. We indicate the share of individuals that hold the cooperation norm by $\phi \in [0,1]$. If ϕ is large, many individuals believe that cooperating is the only morally right thing to do, and we say it is a strong social norm.

Throughout the analysis, we assume that norms are equally distributed in all sub-populations of preference types. Hence, social norm ϕ indicates the proportion of norm holders among all individuals in society as well as among all individuals who hold a particular preference type $\theta \in \text{supp}(\lambda)$. This is a simplifying assumption, further motivated by (1) norm internalization being independent of an individual's preferences (see Section 4.1) and (2) ϕ corresponding to the expected share of norm holders among all individuals who hold the same preference type if, in addition to some mild conditions, differences in the shares of norms holders across preference types arise only due to random occurrences.

Individual i's material payoff derives from her action a^i and the cooperation level ψ according to the payoff-function $m : \mathcal{A} \times [0,1] \to \mathbb{R}$. Throughout, we assume $m(a^i, \psi)$ is continuous and differentiable in its second argument.[6] To capture the nature of public goods games, we assume that cooperating is relatively costly. For simplification purposes, we assume that cooperating becomes relatively more costly in the share of others who cooperate.[7] Appendix A.1 discusses consequences of relaxing this assumption.

Definition 1 *(Material payoff).* $m(a^i, \psi)$ s.t. $\forall \psi \in [0,1]$:

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

Throughout the analysis, we refer to $\Delta m(\psi)$ as the costs of cooperation.

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

Definition 2 *(Self-approval).* $p(a^i, n^i) = (a^i - 1)n^i$.

Social approval captures how i is perceived by her peers and derives from three separate components. First, individuals are subject to social disapproval for social norm violation $(1 - a^i)v(\phi)$. This social disapproval arises as a consequence of acting inappropriately in the eyes of the public. It requires the social norm to be present and increases in its strength, $v(0) = 0$ and $\frac{\mathrm{d}v(\phi)}{\mathrm{d}\phi} > 0$. Second, individuals are subject to social disapproval for non-conformity $k(|n^i - \phi|)$. The more individuals in society hold moral views conflicting with those of i, the greater i's disapproval for non-conformity. Thus, social disapproval for non-conformity increases in the distance between i's personal and the social norm. If i's personal norm coincides with the social one, she experiences no such disapproval. Formally, $\frac{\mathrm{d}k(|n^i-\phi|)}{\mathrm{d}|n^i-\phi|} > 0$ and $k(0) = 0$. Third, individuals are subject to social disapproval for hypocrisy $(1 - a^i)n^i h$, where $h > 0$. An individual i is perceived as a hypocrite if she does not cooperate despite holding the cooperation norm. Social disapproval coincides with negative social approval. So, we write the following.

Definition 3 *(Social approval).* $s(a^i, n^i, \phi) = (a^i - 1)[v(\phi) + n^i h] - k(|n^i - \phi|)$.

[3] In principle, the set of preference types coincides with all possible elements of $\mathbb{R}^2_{\geq 0}$. However, this may create problems regarding the traceability of the evolutionary model, so we adopt this simplifying assumption. Other than for illustrative purposes, we do not impose any restrictions on which preference types exist or may occur by allowing Θ to be arbitrarily large.

[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 benefits an individual. Bašić and Quercia (2022) provide evidence that truth-telling is motivated by social concerns.

[6] 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 is the expected action of a randomly chosen individual.

[7] Larger cooperation levels may require more effort or decrease marginal benefits from the public good.

Table 1
Overview of variables and functions.

Symbol	Description
Societal Framework	
i	individual index
m	material payoff
Δm	costs of cooperation
p	self-approval
s ($\tilde{s}$)	social approval (expressed)
v ($\tilde{v}$)	social disapproval for social norm violation (expressed)
h ($\tilde{h}$)	social disapproval for hypocrisy (expressed)
k ($\tilde{k}$)	social disapproval for non-conformity (expressed)
Δk ($\Delta\tilde{k}$)	difference in social disapproval for non-conformity (expressed)
δ	gossip
Behavior	
a	action
ψ	cooperation share
σ	behavioral distribution
$\sigma_{n,\theta}$	cooperation share of i s.t. $\theta^i = \theta \wedge n^i = n$
σ_n	cooperation share of i s.t. $n^i = n$
u	utility
Σ^*	set of NE

Symbol	Description
Norms	
n	personal norm
ϕ	social norm
c	cultural fitness
C_n	average cultural fitness of i s.t. $n^i = n$
γ	weight of social approval on cultural fitness
$I_p(\lambda)$	potential imperfect-social-norm CE
Approval Preferences	
θ	preference type
θ_s	preference for social approval
θ_p	preference for self-approval
λ	preference distribution
λ_θ	share of i s.t. $\theta^i = \theta$ at λ
supp(λ)	support of λ
b	biological fitness
B_θ	average biological fitness of i s.t. $\theta^i = \theta$
B_λ	average biological fitness of distribution λ
ρ	weight of social approval on biological fitness
θ^d	the dominant preference type
λ^d	preference distribution s.t. $\lambda_{\theta^d} = 1$

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.[8]

Throughout the analysis, we assume that as a consequence of individuals engaging in information sharing (gossip) about their peers' behavior and personal norms, actual social disapproval exceeds the social disapproval expressed towards an individual. We indicate the degree of gossip in society by $\delta \in \mathbb{R}_{\geq 0}$.[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 for social norm violation and non-conformity that is directly communicated to an individual as $\tilde{v}(\phi)$ and $\tilde{k}(|n^i - \phi|)$ respectively. Gossip among peers proportionally increases social disapproval for social norm violation and non-conformity. Hence, we write $v(\phi) = \tilde{v}(\phi)(1+\delta)$ and $k(|n^i - \phi|) = \tilde{k}(|n^i - \phi|)(1+\delta)$. Throughout, $\tilde{v}(\phi)$ and $\tilde{k}(|n^i - \phi|)$ are continuous and differentiable.

The expression of social disapproval for hypocrisy requires observers of i's action to be aware of her respective personal norm and vice versa. Either the observers have previously observed it, or others in society have 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 drawing conclusions therefrom.[10] Thus, social disapproval for hypocrisy is disproportionately greater than the expressed one, $h > \tilde{h}(1+\delta)$.

Definition 4 *(Expressed social approval).* $\tilde{s}(a^i, n^i, \phi) = (a^i - 1)[\tilde{v}(\phi) + n^i\tilde{h}] - \tilde{k}(|n^i - \phi|)$.

Table 1 summarizes the variables and functions most relevant for following the main text of this paper. It also includes terms that we introduce in the upcoming sections.

3. Behavior

Behavior derives from utility maximization and is always in a Nash equilibrium (NE). The underlying idea is that whilst the individuals' cultural and biological characteristics change over time, they choose their behavior rationally at any point in time. An

[8] 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 who believes her peers disapprove of her experiences negative feelings. By changing the setup accordingly, we alter the underlying story but not the formal analysis.

[9] Engaging in gossip may itself be a public goods game to which the results of this paper apply.

[10] For example, let us consider two observers of i in two separate situations so that one may only observe a^i and the other only n^i. Gossip between these two individuals potentially leads to social disapproval of i if it reveals that i behaves inconsistently with her personal norms. However, none of the two observers previously expressed disapproval for hypocrisy towards i.

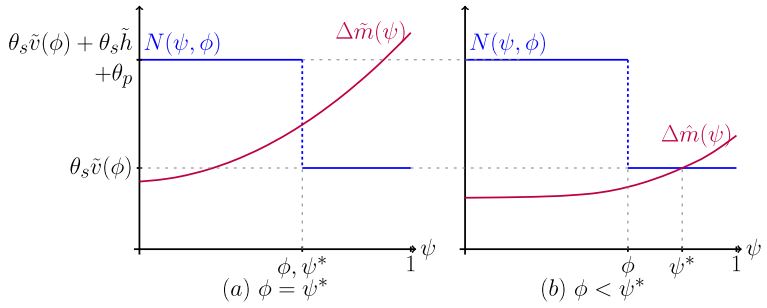

Fig. 1. Equilibrium behavior.

individual i's utility depends on her material payoff $m(a^i,\psi)$, self-approval $p(a^i,n^i)$, and the social approval expressed towards her $\tilde{s}(a^i,n^i,\phi)$. The degree to which these components determine utility depends on her preference type $\theta^i=(\theta^i_s,\theta^i_p)$.

Definition 5 *(Utility).* $u(a^i,n^i,\psi,\phi,\theta^i)=m(a^i,\psi)+\theta^i_s\tilde{s}(a^i,n^i,\phi)+\theta^i_p p(a^i,n^i)$.

Consider any individual with personal norm n and preference type θ. We can inspect her incentives to cooperate by comparing the utility of both actions: $u(1,n,\psi,\phi,\theta)-u(0,n,\psi,\phi,\theta)=\theta_p n+\theta_s\tilde{h}n+\theta_s\tilde{v}(\phi)-\Delta m(\psi)$. An individual who does not hold the cooperation norm, $n=0$, encounters a trade-off between social disapproval for social norm violation $\theta_s\tilde{v}(\phi)$ and costs of cooperation $\Delta m(\psi)$. Norm holders additionally face social disapproval for hypocrisy $\theta_s\tilde{h}$ and self-disapproval θ_p when defecting. Social disapproval for non-conformity does not impact behavioral incentives.

Let $\sigma_{n,\theta}$ be the share of individuals with personal norm n and preference type θ that cooperate. Moreover, $\sigma_n=\sum_{\theta\in\text{supp}(\lambda)}\lambda_\theta\sigma_{n,\theta}$ indicates the share of cooperators among individuals with personal norm n. Hence, we can write the cooperation level as $\psi=\phi\sigma_1+(1-\phi)\sigma_0$. The vector σ consists of all vectors $(\sigma_{1,\theta},\sigma_{0,\theta})$ and, thus, describes the complete distribution of behavior. The following proposition constitutes the main result of this section.

Proposition 1 *(Equilibrium behavior). For each $\lambda\in[0,1]^{|\Theta|}$ and $\phi\in[0,1]$, there is a connected and non-empty set Σ^* s.t.*

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

Proof. Proposition 1 follows from Lemmas 4 and 6 in Appendix B.1.

The above states that all NE at any social norm ϕ and preference distribution λ form a connected and non-empty set Σ^*. Moreover, all NE yield the same cooperation share ψ^*. At times, we may write the set of NE and the equilibrium cooperation level explicitly as functions of the social norm ϕ and preference distribution λ: $\Sigma^*(\phi,\lambda)$ and $\psi^*(\phi,\lambda)$.[11] For ease of readability, the discussion sometimes refers to one specific NE, although Σ^* is not necessarily a singleton. This NE then serves as a representative for all elements in Σ^*, and we write it as σ^*.

To illustrate the intuition behind the above result, we start by discussing a society that is homogeneous regarding approval preferences, $\text{supp}(\lambda)=\{\theta\}$. In this case, $\Sigma^*(\phi,\lambda)$ is a singleton (see Lemma 11 in Appendix B.1). Consider the graphical illustration in Fig. 1. The costs of cooperation at any cooperation level ψ are $\Delta m(\psi)$. By Definition 1, $\Delta m(\psi)$ is strictly increasing. Next, consider the function $N(\psi,\phi)$ that sorts all individuals' social and self-approval gains from cooperation in descending order. The first ϕ individuals hold the cooperation norm and thus avoid social disapproval for social norm violation $\theta_s\tilde{v}(\phi)$, social disapproval for hypocrisy $\theta_s\tilde{h}$, and self-disapproval θ_p when cooperating. The remaining $1-\phi$ individuals do not hold the cooperation norm and thus only avoid social disapproval for social norm violation $\theta_s\tilde{v}(\phi)$. $N(\psi,\phi)$ is a decreasing function by construction.

If at some cooperation level ψ, $N(\psi,\phi)$ lies above $\Delta m(\psi)$, an individual currently not cooperating prefers to do so. Similarly, if $N(\psi,\phi)$ lies below $\Delta m(\psi)$, an individual who cooperates prefers to defect. The NE corresponds to the unique intersection of both curves or, if no intersection exists, to one of the boundary points. Fig. 1a shows an example in which all norm holders strictly prefer to cooperate, and all norm non-holders strictly prefer to defect. Fig. 1b provides an example of some non-holders being sufficiently motivated to cooperate. The costs of cooperation equal social disapproval for social norm violation in a manner that all norm non-holders are indifferent.

When preferences are heterogeneous, we can also sort all individuals according to their utility gains from cooperating, yielding a decreasing function. The unique equilibrium share of cooperators ψ^* is at this function's intersection with the costs of cooperation $\Delta m(\psi)$. However, the set of all NE is no longer necessarily a singleton but possibly a connected set. This may occur if at least two

[11] Note that we can show that Σ^* would be asymptotically stable if behavior was to follow an evolutionary process of pairwise comparison dynamics (see Lemma 4 in Appendix B.1), providing additional reasoning as to why we expect society to reach an element in this set.

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 between both behavioral routines, $\hat{\theta}_s \tilde{v}(\phi) + \hat{n}\hat{\theta}_s \tilde{h} + \hat{n}\hat{\theta}_p = \Delta m(\psi^*) = \check{\theta}_s \tilde{v}(\phi) + \check{n}\check{\theta}_s \tilde{h} + \check{n}\check{\theta}_p$. If so, there may exist infinitely many NE σ^* that all exhibit varying sub-group cooperation levels $\sigma^*_{\hat{\theta},\hat{n}}$ and $\sigma^*_{\check{\theta},\check{n}}$, but the same total cooperation level ψ^*.

We can use the above results to examine how different variables and functions influence the equilibrium cooperation level ψ^*. To do so, we focus on a society with homogeneous approval preferences since, in the long run (when preferences are endogenous), all individuals behave as if their preferences were homogeneous. Hence, to analyze the effect of different variables and functions on the equilibrium outcome, it suffices to analyze how they affect equilibrium behavior under the homogeneous preference distribution that society mimics.

Proposition 2 *(Comparative results for equilibrium behavior). Consider any $\lambda \in [0,1]^{|\Theta|}$ s.t. $\text{supp}(\lambda) = \{\theta\}$. The equilibrium cooperation share ψ^* is (weakly) greater for greater negative costs of cooperation $-\Delta m(\cdot)$, preference for self-approval θ_p, preference for social approval θ_s, social norm ϕ, expressed disapproval for social norm violation $\tilde{v}(\cdot)$, and expressed social disapproval for hypocrisy $\tilde{h}$.*

Proof. Lemma 8 and Proposition 16 in Appendix B.1 constitute the formal equivalent to the above.

Consider Fig. 1 and note that an increase in $-\Delta m(\psi)$ corresponds to a decrease in the costs of cooperation $\Delta m(\psi)$. Such a decrease shifts the respective curve downwards and its intersection with $N(\psi, \phi)$ to the right. The equilibrium cooperation share must increase. An increase in the social norm ϕ implies that more individuals hold the cooperation norm. Hence, social disapproval for social norm violation also increases. Graphically, both horizontal segments of $N(\psi, \phi)$ shift upwards, and their vertical connection moves to the right. Consequently, the equilibrium level of cooperation must rise. By similar reasoning, increases in θ_s and $\tilde{v}(\phi)$ move both segments, and increases in θ_p and $\tilde{h}$ move the left segment of $N(\psi, \phi)$ upwards, which (weakly) increases the equilibrium cooperation share ψ^*.

4. Norms

4.1. Evolutionary framework

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 factors (e.g., social reputation, respect). Thus, material payoff and social approval co-determine the cultural fitness that drives the evolution of norms.

Definition 6 *(Cultural fitness).* $c(a^i, n^i, \psi, \phi) = m(a^i, \psi) + \gamma s(a^i, n^i, \phi)$, where $0 < \gamma$ is the weight of social approval on cultural fitness.

Following the existing literature, we assume that 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 (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), in turn increasing the impact on the opinion formation process of others (Bowles and Gintis, 1998; Gintis, 2003b). Access to certain social networks is often denied if an individual is subject to social disapproval (Cinyabuguma et al., 2005; Traxler and Spichtig, 2011). Since norms evolve based on learning through socialization, and there is no evident systematic relationship to an individual's preferences, we assume that norm internalization occurs independently of an individual's approval preferences. Formally, we can best describe the cultural evolution of norms using imitative dynamics (see Sandholm, 2010). Therefore, we employ the well-studied replicator dynamics.[12]

Definition 7 *(Norm dynamics).*

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

where $\Delta k(\phi) = k(|0-\phi|) - k(|1-\phi|)$ is the difference in social disapproval for non-conformity between norm non-holders and holders, and $C_n(\sigma, \phi) = \sigma_n c(1, n, \psi, \phi) + (1-\sigma_n)c(0, n, \psi, \phi)$ is the average cultural fitness of all individuals with personal norm $n \in \{0, 1\}$.

Throughout, we refer to a rest point of norm dynamics as a *cultural equilibrium (CE)* and may indicate it by ϕ^*. A *stable* CE is one for which the social norm always remains close to it when starting sufficiently close to the CE. The CE is *asymptotically stable* if the social norm converges to it.

[12] Following Sandholm (2010), we can show that the results of this paper also hold for a variety of other population dynamics.

4.2. Equilibrium analysis

We continue by discussing the results most relevant to the paper's further analysis and findings.[13] To this end, we investigate and discuss the existence and stability of different CE for an exogenous preference distribution λ.

4.2.1. No-social-norm cultural equilibrium

The first CE corresponds to the absent social norm $\phi^* = 0$, where no individual holds the cooperation norm. All individuals have neither personal nor social incentives to cooperate and, thus, defect, $\psi^*(0, \lambda) = 0$. Such a CE always exists and is always asymptotically stable.

Proposition 3 *(Asymptotically stable no-social-norm CE). For any $\lambda \in [0,1]^{|\Theta|}$, $\phi = 0$ is an asymptotically stable CE.*

Proof. See Appendix B.2.

Starting from $\phi^* = 0$, assume that a small group of norm holders appears. As the new social norm ϕ remains near zero, the norm holders are subject to high social disapproval for non-conformity. Moreover, internalizing the cooperation norm may induce them to either change their action to cooperation or keep defecting. In the former case, the norm holders incur material costs but avoid social disapproval for social norm violation. However, the avoided social disapproval minimally impacts the difference in average cultural fitness as the social norm ϕ is close to zero. In the latter case, everyone obtains the same material payoff and social disapproval for 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 $\phi^* = 0$.

4.2.2. Perfect-social-norm cultural equilibrium

Next, we discuss the perfect-social-norm CE $\phi^* = 1$, where all individuals hold the cooperation norm. The perfect-social-norm CE always exists.

Proposition 4 *(Asymptotically stable perfect-social-norm CE). For any $\lambda \in [0,1]^{|\Theta|}$, $\phi = 1$ is an asymptotically stable CE if*

1. $\psi^*(1,\lambda)(\gamma \upsilon(1) - \Delta m(\psi^*(1,\lambda))) - \gamma(1 - \psi^*(1,\lambda))h + \gamma \Delta k(1) > 0$ *and*
2. (a) $\theta_s \tilde{\upsilon}(1) < \Delta m(\psi^*(1,\lambda))\ \forall \theta \in \operatorname{supp}(\lambda)$ *or*
   (b) $\Delta k(1) > (1 - \psi^*(1,\lambda))h$.

Proof. See Appendix B.2.

Proposition 4 states that the perfect-social-norm CE is asymptotically stable if (1) the norm holders, corresponding to all individuals in society, obtain greater cultural fitness on average than a hypothetical group of norm non-holders would if they were not cooperating and (2) either (a) any individual would prefer to defect if she was not holding the cooperation norm or (b) the difference in social disapproval for non-conformity outweighs average social disapproval for hypocrisy.

At the perfect social norm $\phi^* = 1$, society consists of only norm holders, implying that the cooperation share $\psi^*(1,\lambda)$ coincides with σ_1^*. Consider a small cultural mutation to social norm $\phi < 1$. Since the post-mutation social norm ϕ is close to one, the behavioral incentives for norm holders change only slightly. Consequently, the norm-holder cooperation share σ_1^* and the total cooperation share $\psi^*(\phi,\lambda)$ remain close to the pre-mutation cooperation share $\psi^*(1,\lambda)$. Hence, after the cultural mutation, it holds that the norm holders obtain greater cultural fitness on average than the norm non-holders if all norm non-holders defect, $\sigma_1^*(\gamma \upsilon(\phi) - \Delta m(\psi^*)) + \gamma(1 - \sigma_1^*)h + \Delta k(\phi) > 0$.

Suppose the individuals who abandon the cooperation norm due to cultural mutation are no longer sufficiently motivated to cooperate, $\sigma_0^* = 0$. In that case, the average cultural fitness of norm holders exceeds that of norm non-holders at the new social norm ϕ, and cultural evolution reinstates the perfect social norm $\phi^* = 1$. This is the case if Condition 2a holds: If all individuals would strictly prefer to defect if they were not holding the cooperation norm at the perfect social norm $\phi^* = 1$, $\theta_s \tilde{\upsilon}(\phi) < \Delta m(\psi^*(1,\lambda))$, they prefer to defect when not holding the cooperation norm at any social norm ϕ close to $\phi^* = 1$, $\theta_s \tilde{\upsilon}(\phi) < \Delta m(\psi^*(\phi,\lambda))$.

Alternatively, suppose that some individuals who abandon the cooperation norm prefer to cooperate, $\sigma_0^* > 0$. As a consequence, differences in average social disapproval for social norm violation and material costs of both norm holder populations become less pronounced when compared to the case in which all norm non-holders defect, $(\sigma_1^* - \sigma_0^*)(\gamma \upsilon(\phi) - \Delta m(\psi^*)) < \sigma_1^*(\gamma \upsilon(\phi) - \Delta m(\psi^*))$. This increases the relative impact of differences in social disapproval for non-conformity $\Delta k(\phi)$ and hypocrisy $(1 - \sigma_1^*)h$ on differences in cultural fitness. Condition 2b ensures that after the cultural mutation to social norm ϕ, the difference in social disapproval for non-conformity outweighs average social disapproval for hypocrisy, $\Delta k(\phi) > (1 - \sigma_1^*)h$. Hence, norm holders obtain more cultural fitness than norm non-holders on average, and cultural evolution reinstates the perfect social norm $\phi^* = 1$.

[13] Appendix A.2 briefly discusses some additional CE that may exist.

The above highlights the importance of social disapproval for non-conformity in stabilizing a perfect social norm. Similarly, the proposition illustrates that social disapproval for hypocrisy may destabilize the perfect social norm as it negatively affects Conditions 1 and 2b. The underlying reason is that only the norm holders can behave hypocritically and thus hold an evolutionary disadvantage.

Proposition 5 *(Robustness of an asymptotically stable perfect-social-norm CE). Consider any $\lambda \in [0,1]^{|\Theta|}$ s.t. there exists a CE of Proposition 4. There is a neighborhood U of λ s.t. $\phi = 1$ is an asymptotically stable CE at any $\hat{\lambda} \in U$.*

Proof. See Appendix B.2.

Proposition 5 establishes that if the perfect social norm is an asymptotically stable CE of Proposition 4 at preference distribution λ, then it is also an asymptotically stable CE at any other preference distribution $\hat{\lambda}$ that is sufficiently close to λ. The underlying intuition is that at preference distribution λ and any social norm ϕ in the neighborhood of $\phi^* = 1$, the norm holders obtain strictly greater cultural fitness than the norm non-holders, $(\sigma_1^* - \sigma_0^*)(\gamma \upsilon(\phi) - \Delta m(\phi)) + \gamma(1 - \sigma_1^*)h + \Delta k(\phi) > 0$. A small change in preferences from λ to $\hat{\lambda}$ only alters norm population behavior (σ_1^*, σ_0^*) at any social norm ϕ slightly. Hence, the inequality above still holds at $\hat{\lambda}$. Thus, the norm holders still obtain greater cultural fitness than the norm non-holders at preference distribution $\hat{\lambda}$ and any social norm ϕ in the neighborhood of $\phi^* = 1$. The perfect social norm is an asymptotically stable CE at preference distribution $\hat{\lambda}$.

Finally, we analyze how different variables and functions may relate to the existence of an asymptotically stable CE of Proposition 4. We focus on the special case of homogeneous preferences, $\text{supp}(\lambda) = \{\theta\}$, where approval preferences are relatively small as compared to the weight of social approval on cultural fitness, $\theta_s < \gamma(1+\delta)$ and $\theta_p < \gamma(h - (1+\delta)\tilde{h})$. We do so since this constitutes the relevant case for further analysis in Section 5. The results are summarized in the following proposition.

Proposition 6 *(Comparative results for an asymptotically stable perfect-social-norm CE). Consider any specification of the model with $\lambda \in [0,1]^{|\Theta|}$ s.t. $\text{supp}(\lambda) = \{\theta\}$ and $(\theta_s, \theta_p) < (\gamma(1+\delta), \gamma(h - (1+\delta)\tilde{h}))$.*

1. *The CE of Proposition 4 more likely exists if the weight of social approval on cultural fitness γ, social disapproval for social norm violation $\upsilon(\cdot)$, the difference in social disapproval for non-conformity $\Delta k(\cdot)$, and negative costs of cooperation $-\Delta m(\cdot)$ are large.*
2. *There exists a CE of Proposition 4 if*
   (a) *social disapproval for hypocrisy h and/or costs of cooperation $\Delta m(\cdot)$ are small, and/or*
   (b) *the difference in social disapproval for non-conformity $\Delta k(\cdot)$ and/or social norm violation $\upsilon(\cdot)$ are large.*
3. *A small increase in ψ^*, ceteris paribus, ambiguously affects the conditions of Proposition 4.*

Proof. Proposition 17 in Appendix B.2 constitutes the formal equivalent to the above.

Regarding the existence of an asymptotically stable perfect-social-norm CE of Proposition 4, the first two statements of Proposition 6 ascribe a negative role to social disapproval for hypocrisy h and the costs of cooperation $\Delta m(\cdot)$, while ascribing a positive role to the weight of social approval on cultural fitness γ, social disapproval for social norm violation $\upsilon(\cdot)$, and the difference in social disapproval for non-conformity $\Delta k(\cdot)$. The perfect social norm is an asymptotically stable CE if and only if the average cultural fitness of the norm holders is greater than that of norm non-holders at social norms ϕ close to the perfect one $\phi^* = 1$. Hence, we can illustrate why the above proposition holds by discussing how the different variables and functions impact differences in cultural fitness.

For this purpose, recall that individuals who hold the cooperation norm always have more incentives to cooperate, implying that their share of cooperators must always be greater than the share among norm non-holders, $\sigma_1^* \geq \sigma_0^*$. Consequently, greater social disapproval for social norm violation $\upsilon(\cdot)$ impacts norm holders' average cultural fitness less than that of norm non-holders. Similarly, if the cooperation costs are small, cooperating norm holders forego less material payoff, positively impacting their cultural fitness. For the perfect social norm to be asymptotically stable, it must hold that the norm holders are subject to more social approval than the norm non-holders in some neighborhood of the perfect social norm. Otherwise, the norm non-holders would, on average, obtain higher material payoff (since fewer of them cooperate) and social approval. If the norm holders' average social approval is larger, then differences in cultural fitness increase in the weight of social approval on cultural fitness γ in their favor. Furthermore, a greater difference in social disapproval for non-conformity $\Delta k(\cdot)$ positively impacts the cultural fitness of norm holders at any social norm close to the perfect one. Lastly, social disapproval for hypocrisy h may hinder asymptotic stability of a perfect-social-norm CE since only the carriers of the cooperation norm can be subject to it. Hence, they hold an evolutionary disadvantage, possibly hindering the spread of the cooperation norm. Note that for social disapproval for hypocrisy to prevent the perfect social norm from being asymptotically stable, hypocrisy must occur, $\sigma_1^* < 1$.

The third statement of Proposition 6 holds due to two countervailing effects on differences in cultural fitness stemming from an increase in the cooperation share ψ^*. On the one hand, increasing the cooperation share ψ^* reduces average social disapproval for hypocrisy $(1 - \psi^*)h$, increasing the norm holders' average cultural fitness. On the other hand, an increase in the cooperation share ψ^* raises the costs of cooperation $\Delta m(\psi^*)$, negatively impacting the cultural fitness of the cooperating norm holders. Which effect dominates depends on how responsive the cooperation costs $\Delta m(\cdot)$ are with respect to ψ^*. Generally, a larger responsiveness corresponds to a greater cost increase due to an increased cooperation share ψ^*, thus favoring domination of the latter effect.

Note that the third statement of Proposition 6 implies that changes in the preferences for approval θ_s and θ_p as well as expressed social disapproval for hypocrisy $\tilde{h}$ and social norm violation $\tilde{\upsilon}(\cdot)$ affect the existence of an asymptotically stable perfect-social-norm CE

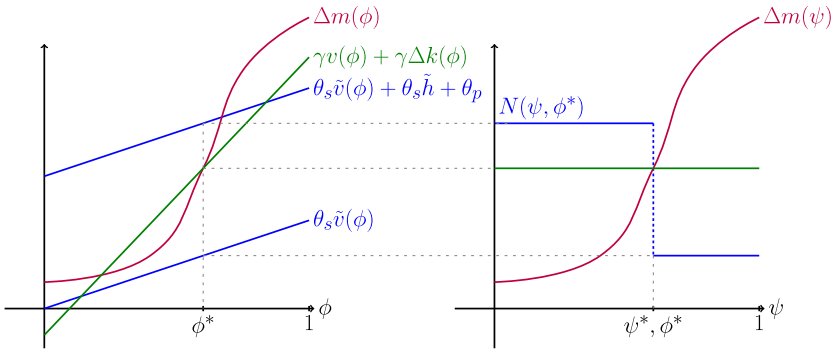

Fig. 2. Imperfect-social-norm CE.

indirectly through the cooperation share (recall Proposition 2 in Section 3). Consequently, changes in social disapproval for hypocrisy and social norm violation affect it directly through their actual values h and $v(\cdot)$ and indirectly through their expressed counterparts $\tilde{h}$ and $\tilde{v}(\cdot)$. Although these indirect effects have not been accounted for in the discussion of Proposition 6, the formal results still hold when doing so.

4.2.3. *Imperfect-social-norm cultural equilibrium*

The following proposition describes an asymptotically stable CE of an imperfect social norm $\phi^* \in (0,1)$.

Proposition 7 *(Asymptotically stable imperfect-social-norm CE). For any $\lambda \in [0,1]^{|\Theta|}$, $\phi \in (0,1)$ is an asymptotically stable CE if*

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

Proof. See Appendix B.2.

Condition 1 implies that at social norm ϕ^* all norm non-holders strictly prefer to defect and all norm holders strictly prefer to cooperate, implying norm population behavior is $(\sigma_1^*, \sigma_0^*) = (1,0)$ for all social norms in the neighborhood of ϕ^*. Condition 2 ensures that social norm ϕ^* is a CE, while Condition 3 implies it is asymptotically stable.

Fig. 2 presents a graphical illustration of the proposition for the case of homogeneous preferences, $\text{supp}(\lambda) = \{\theta\}$. The intuition easily carries over to the heterogeneous preferences case. At the respective imperfect social norm and some neighborhood around it, individuals cooperate if and only if they hold the cooperation norm, $(\sigma_1^*, \sigma_0^*) = (1,0)$. For norm dynamics 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 cooperation $\Delta m(\phi^*)$ equal the differences in social disapproval for social norm violation and non-conformity on cultural fitness $\gamma v(\phi^*) + \gamma \Delta k(\phi^*)$. 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 CE ϕ^*. This applies if an increase (decrease) in ϕ^* increases (decreases) the costs of cooperation more than the difference in social approval. Graphically, $\Delta m(\phi)$ intersects $\gamma v(\phi) + \gamma \Delta k(\phi)$ from below at ϕ^*. Hence, asymptotic stability of the CE requires material payoff to be relatively more responsive to changes in the cooperation share than social disapproval is to changes in the social norm.

Proposition 8 *(Robustness of an asymptotically stable imperfect-social-norm CE). Consider any $\lambda \in [0,1]^{|\Theta|}$ s.t. there exists an imperfect-social-norm CE $\phi^* \in (0,1)$ of Proposition 7. For all $\epsilon > 0$, there is a neighborhood U of λ s.t. at any $\hat{\lambda} \in U$, there exists an asymptotically stable set $\hat{\Phi} \subset (\phi^* - \epsilon, \phi^* + \epsilon)$.*

Proof. Follows from Proposition 18 in Appendix B.2.

Proposition 8 implies that if there exists an imperfect-social-norm CE ϕ^* of Proposition 7 at preference distribution λ, then for all preference distributions $\hat{\lambda}$ close to λ, there exists a CE $\hat{\phi}^*$ close to ϕ^* that society coordinates into at preference distribution $\hat{\lambda}$ and social norm ϕ^*. Fig. 3 illustrates the underlying intuition of this result graphically. At preference distribution λ, the right intersection of the two solid lines constitutes the asymptotically stable CE ϕ^*. 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 ϕ^* 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 ϕ^* is $\sigma_1^* = 1 - \epsilon_1$, $\sigma_0^* = \epsilon_0$, and $\psi^* = \phi^*(1-\epsilon_1) + (1-\phi^*)\epsilon_0$. Substituting this into Definition 7 yields that norm dynamics are at rest if $(1-\epsilon_1-\epsilon_0)(\gamma v(\phi) - \Delta m(\phi(1-\epsilon_1) + (1-\phi)\epsilon_0)) - \gamma \epsilon_1 h + \gamma \Delta k(\phi) = 0$. The intersection of the two dotted lines at $\hat{\phi}^*$ in Fig. 3 represents such a rest point. It is asymptotically stable since $\Delta m(\phi(1-\epsilon_1) + (1-\phi)\epsilon_0)$

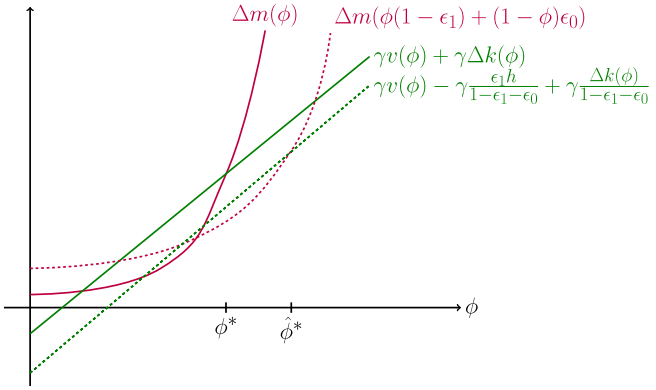

Fig. 3. Robustness of an imperfect-social-norm CE.

intersects $\gamma v(\phi)+\gamma\frac{\epsilon_1 h}{1-\epsilon_1-\epsilon_0}+\gamma\frac{\Delta k(\phi)}{1-\epsilon_1-\epsilon_0}$ from below. The closer the two preference distributions $\hat{\lambda}$ and λ, the closer the dotted lines are to the solid lines and, consequently, $\hat{\phi}^*$ is to ϕ^*.

We continue by analyzing how different variables and functions may affect the potential existence as well as an existing imperfect-social-norm CE of Proposition 7. For this purpose, we again focus on the case of homogeneous preferences, $\text{supp}(\lambda)=\{\theta\}$, where approval preferences are relatively small as compared to the weight of social approval on cultural fitness, $\theta_s<\gamma(1+\delta)$ and $\theta_p<\gamma(h-(1+\delta)\tilde{h})$.

Conditions 1 and 2 of Proposition 7 imply that an imperfect social norm ϕ can only be a CE of Proposition 7 if the differences in social disapproval for social norm violation and non-conformity on cultural fitness exceed the norm-based cooperation incentives of norm non-holders but fall short of the norm-based cooperation incentives of norm holders, $\theta_s\tilde{v}(\phi)<\gamma v(\phi)+\gamma\Delta k(\phi)<\theta_s\tilde{v}(\phi)+\theta_s\tilde{h}+\theta_p$. Graphically, this corresponds to the green curve being between the two blue curves in Fig. 2. We refer to all social norms ϕ satisfying this condition as the set of potential imperfect-social-norm CE $I_p(\lambda)$.

Lemma 1 *(Potential imperfect-social-norm CE). Consider any $\lambda\in[0,1]^{|\Theta|}$ s.t. $\text{supp}(\lambda)=\{\theta\}$ and $(\theta_s,\theta_p)<(\gamma(1+\delta),\gamma(h-(1+\delta)\tilde{h}))$. Let $I_p(\lambda)=\{\phi\in[0,1]:\theta_s\tilde{v}(\phi)<\gamma v(\phi)+\gamma\Delta k(\phi)<\theta_s\tilde{v}(\phi)+\theta_s\tilde{h}+\theta_p\}$.*

1. *$I_p(\lambda)$ increases in $\tilde{h}$ and θ_p and*
2. *$\theta_s\tilde{h}+\theta_p>0\Leftrightarrow I_p(\lambda)\neq\emptyset$.*

Proof. See Appendix B.2.

Lemma 1 states that the set of potential imperfect-social-norm CE $I_p(\lambda)$ (1) increases in expressed social disapproval for hypocrisy $\tilde{h}$ and the preference for self-approval θ_p and (2) is non-empty if and only if expressed social disapproval for hypocrisy and self-approval affect utility, $\theta_s\tilde{h}+\theta_p>0$. The underlying reason is that expressed social disapproval for hypocrisy $\tilde{h}$ and the preference for self-approval θ_p are responsible for the difference in cooperation incentives of norm holders and non-holders. Graphically, an increase in $\tilde{h}$ and θ_p shifts the norm holders cooperation incentives $\theta_s\tilde{v}(\phi)+\theta_s\tilde{h}+\theta_p$ upwards, whilst leaving the norm non-holders cooperation incentives $\theta_s\tilde{v}(\phi)$ unaffected in Fig. 2. The set of potential imperfect-social-norm CE $I_p(\lambda)$ increases. If expressed social disapproval for hypocrisy and self-approval were not to affect utility, $\theta_s\tilde{h}+\theta_p=0$, then norm holders and non-holders would have the same incentives to cooperate at all social norms ϕ. Both blue lines in Fig. 2 coincide, and Condition 1 of Proposition 7 can never be satisfied. The set of potential imperfect-social-norm CE $I_p(\lambda)$ is the empty set.

Note that for any social norm ϕ in the set of potential imperfect-social-norm CE $I_p(\lambda)$, there exist cost curves $\Delta m(\cdot)$ that induce ϕ to be an asymptotically stable CE, namely those cost curves that yield Conditions 2 and 3 of Proposition 7 to hold. Hence, the above highlights the positive role of expressed social disapproval for hypocrisy $\tilde{h}$ and self-approval θ_p for the potential existence of an asymptotically stable imperfect-social-norm CE.

Finally, we investigate the effect of changes in different variables and functions on an existing imperfect-social-norm CE ϕ^* of Proposition 7 and, hence, the corresponding cooperation level $\psi^*=\phi^*$.

Proposition 9 *(Comparative results for an asymptotically stable imperfect-social-norm CE). Consider any $\lambda\in[0,1]^{|\Theta|}$ s.t. an imperfect-social-norm CE $\phi^*\in(0,1)$ of Proposition 7 exists.*

1. *For some larger weight of social approval on cultural fitness γ, social disapproval for social norm violation $v(\cdot)$, or negative costs of cooperation $-\Delta m(\cdot)$, there exists an asymptotically stable set $\hat{\Phi}$ s.t. $\hat{\phi}>\phi^*$ for all $\hat{\phi}\in\hat{\Phi}$.*
2. *Suppose $\phi^*>\frac{1}{2}$ ($\phi^*<\frac{1}{2}$). A change in social disapproval for non-conformity $k(\cdot)$ corresponding to a larger absolute difference $|\Delta k(\cdot)|$ leads to the existence of an asymptotically stable set $\hat{\Phi}$ s.t. $\hat{\phi}>\phi^*$ ($\hat{\phi}<\phi^*$) for all $\hat{\phi}\in\hat{\Phi}$.*

Proof. Proposition 19 in Appendix B.2 constitutes the formal equivalent to the above.

First, the proposition implies that slight increases in the weight of social approval on cultural fitness γ, social disapproval for social norm violation $v(\cdot)$, and negative costs of cooperation $-\Delta m(\cdot)$ induce an imperfect-social-norm CE ϕ^* to rise. The underlying reason is that the described changes impact differences in cultural fitness in favor of the norm holders, strengthening the social norm. Graphically, an increase in social disapproval for social norm violation $v(\cdot)$ shifts $\gamma v(\phi) + \gamma \Delta k(\phi)$ upwards in Fig. 2. Similarly, an increase in the weight γ rotates $\gamma v(\phi) + \gamma \Delta k(\phi)$ counterclockwise around its x-intercept. An increase in $-\Delta m(\cdot)$ corresponds to a downward shift of the cost curve $\Delta m(\psi)$. These movements of the two curves imply that their intersection, the CE, moves to the right. Hence, the CE strengthens.

Second, the proposition makes some inferences regarding social disapproval for non-conformity $k(\cdot)$. The effect of a change in social disapproval for non-conformity depends on the relative strength of the social norm. In particular, if the difference in social disapproval for non-conformity becomes more pronounced, a relatively strong CE further strengthens, whereas a relatively weak CE further weakens. Graphically, increasing $|\Delta k(\cdot)|$ rotates $\gamma v(\phi) + \gamma \Delta k(\phi)$ in Fig. 2 counterclockwise around it's point at $\frac{1}{2}$ (recall that at this point $\Delta k(\phi) = 0$). This shifts its intersection with $\Delta m(\phi)$, the CE, to the right or left, depending on whether $\phi^* > \frac{1}{2}$ or $\phi^* < \frac{1}{2}$ respectively. Hence, social disapproval for non-conformity supports the persistence of relatively strong social norms $\phi^* > \frac{1}{2}$ even if other forces favor a weakening. By similar reasoning, conformity concerns can trap a relatively weak social norm $\phi^* < \frac{1}{2}$, despite other forces supporting a further spread.

Small changes in approval preferences θ and expressed social disapproval for hypocrisy $\tilde{h}$ do not affect the situation. Such changes alter the cooperation incentives of individuals. However, if these alterations are sufficiently small, they do not change equilibrium behavior $(\sigma_1^*, \sigma_0^*) = (1, 0)$ and, hence, the equilibrium cooperation share $\psi^* = \phi$. Consequently, to (marginally) impact the equilibrium cooperation level ψ^* at an imperfect-social-norm CE ϕ^*, we need to target it indirectly through the CE ϕ^*.

5. Approval preferences

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

5.1. Evolutionary framework

Approval preferences evolve through biological reproduction.[14] Like cultural fitness, material payoff and social approval co-determine the fitness that drives biological reproduction. Formally, we write an individual's biological fitness as follows.

Definition 8 *(Biological fitness).* $b(a^i, n^i, \psi, \phi) = m(a^i, \psi) + \rho s(a^i, n^i, \phi)$, where $0 < \rho < \gamma$ is the weight of social approval on biological fitness.

Individuals with relatively high biological fitness have greater access to social and material resources, positively affecting their parenting abilities (Irons, 1979; Geary et al., 2004). This increases their reproductive fitness through greater survival chances of their offspring (Turke, 1989; Buss and Schmitt, 1993; 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 imitative dynamics.

Definition 9 *(Preference dynamics).*

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

where

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

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 at preference distribution λ respectively. In line with the equilibrium notion for norms, we call a rest point of preference dynamics a *biological equilibrium (BE)*. At times, we indicate a BE by λ^*. We write a pair of a preference distribution and social norm as the vector (λ, ϕ) and employ stability notions as introduced in Section 4.

[14] Note that vertical transmission (parents to offspring) of preferences not only occurs through biological reproduction but also cultural socialization. However, to avoid confusion in the text regarding the evolutionary processes of preferences and norms, we refer to all transmission processes from parents to offspring as biological reproduction.

5.2. Equilibrium analysis

This section proceeds with the formal analysis when preference formation is endogenous. First, we introduce a preference type that always induces biological fitness-maximizing behavior. Based on this preference type, we characterize BE, whose dynamic stability we then investigate. In particular, for each CE ϕ^* of Section 4, we investigate whether a BE λ^* and the CE ϕ^* form a stable pair (λ^*, ϕ^*). By focusing on the asymptotically stable CE presented in Section 4, we ensure that (λ^*, ϕ^*) is always stable on the norm dimension, allowing us to focus our discussion on the biological evolution of preferences.

5.2.1. The dominant preference type and biological equilibria

We start by presenting the preference type that always induces biological fitness-maximizing behavior.

Definition 10 *(The dominant preference type).* $\theta^d := (\rho(1+\delta), \rho(h-(1+\delta)\tilde{h}))$.

Let λ^d be the preference distribution for which only preference type θ^d exists. Any individual with preference type θ^d experiences disutility from expressed social disapproval equal to the degree that it is proportionally increased through gossip δ and then impacts biological fitness, $\theta^d_s = \rho(1+\delta)$. The preference for self-approval accounts for social disapproval for hypocrisy that arises from information pooling, $\theta^d_p = \rho(h-(1+\delta)\tilde{h})$. By substituting θ^d_s and θ^d_p into the utility function, it becomes apparent that the utility of an individual with preference type θ^d mimics biological fitness.

Suppose some individuals are of preference type θ^d at preference distribution λ, $\theta^d \in \text{supp}(\lambda)$. Since equilibrium behavior maximizes an individual's utility, equilibrium behavior of individuals with preference type θ^d maximizes their biological fitness (see Lemma 19 in Appendix B.3). Hence, individuals with preference type θ^d always obtain (weakly) greater biological fitness than their peers, $B_{\theta^d}(\sigma^*, \phi) \geq B_\theta(\sigma^*, \phi)\ \forall \theta \in \Theta$. Unless all individuals currently maximize their biological fitness, the preference type θ^d spreads in society, $\dot{\lambda}_{\theta^d} = 0 \Leftrightarrow B_{\theta^d}(\sigma^*, \phi) = B_\theta(\sigma^*, \phi)\ \forall \theta \in \text{supp}(\lambda)$. Now suppose that preference type θ^d is not necessarily present at distribution λ. If all individuals behave as if their preference type was θ^d, they maximize biological fitness, and preference dynamics are at rest. Moreover, if all individuals behave as if their preference type was θ^d, norm population behavior (σ^*_1, σ^*_0) is as if preferences were distributed according to λ^d. Given this reasoning, we infer that any preference distribution λ and CE ϕ^* constitute a stable BE and CE if norm population behavior is as if preferences were distributed according to λ^d. The following lemma captures this result formally.

Lemma 2 *(BE and CE). For any* $\lambda \in [0,1]^{|\Theta|}$ *s.t.* $\phi^* \in [0,1]$ *is a CE,* (λ, ϕ^*) *is a rest point if*

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

Proof. See Appendix B.3.

Note that the second condition requires that if all individuals share the same personal norm, $\phi^* \in \{0,1\}$, the equilibrium cooperation share is as if only the dominant preference type existed, $\psi^*(\phi^*, \lambda) = \psi^*(\phi^*, \lambda^d)$. In that case, actual norm population behavior is fully described by only the existing norm population's behavior, which coincides with the overall equilibrium cooperation share, $\sigma^*_{\phi^*} = \psi^*(\phi^*, \lambda)$. Therefore, Condition 2 is equivalent to stating that for a no- or perfect-social norm CE, $\phi^* \in \{0,1\}$, equilibrium norm population behavior at preference distributions λ and λ^d is alike.

5.2.2. Biological equilibrium and no-social-norm cultural equilibrium

Next, we establish that any preference distribution λ is a stable BE for the no-social-norm CE.

Proposition 10 *(Stable BE and no-social-norm CE). For any* $\lambda \in [0,1]^{|\Theta|}$, $(\lambda, 0)$ *is a stable rest point.*

Proof. See Appendix B.3.

Recall that at any preference distribution λ, the no-social-norm CE $\phi^* = 0$ exists and is asymptotically stable (see Proposition 3 in Section 4.2.1). Moreover, all individuals always defect since they have neither personal nor social incentives to cooperate, $\psi^*(0, \lambda) = 0$. Consequently, all individuals behave alike, and biological evolution is at rest, $\dot{\lambda}_\theta = 0\ \forall \theta \in \Theta$. A biological mutation from preference distribution λ to $\hat{\lambda}$ neither alters the CE $\phi^* = 0$ nor equilibrium behavior $\psi^*(0, \hat{\lambda}) = 0$. The dynamic system remains at rest. Hence, preference distribution λ and the no-social-norm CE indeed correspond to a stable rest point of the dynamic system. Note that the above implies that although always stable, any $(\lambda, 0)$ is never asymptotically stable. Biological mutations across the set of all possible preference distributions, however, never disrupt the situation.

5.2.3. Biological equilibrium and perfect-social-norm cultural equilibrium

Next, we investigate the existence of a stable BE and perfect-social-norm CE.

Proposition 11 *(Stable BE and perfect-social-norm CE). For any $\lambda \in [0,1]^{|\Theta|}$, $(\lambda, 1)$ is a stable rest point if (1) $\psi^*(1,\lambda) = \psi^*(1,\lambda^d)$ and (2) $\phi = 1$ is a CE of Proposition 4 at λ.*

Proof. See Appendix B.3.

Consider any preference distribution λ and suppose the conditions of Proposition 11 hold. Recall that Condition 1 of Proposition 11 implies that λ is a BE. Since Condition 2 implies that the perfect social norm is an asymptotically stable CE at preference distribution λ, $(\lambda, 1)$ constitutes a rest point of the dynamic system.

Consider a small biological mutation to some preference distribution $\hat{\lambda}$. Since the post-mutation preference distribution $\hat{\lambda}$ is close to λ, the perfect social norm remains an asymptotically stable CE (recall Proposition 5 in Section 4.2.2). Hence, the social norm does not change, and everyone holds the cooperation norm, $\phi^* = 1$. If the biological mutation does not alter the cooperation share, $\psi^*(1,\lambda) = \psi^*(1,\lambda^d)$, then all individuals maximize biological fitness at preference distribution $\hat{\lambda}$. The dynamic system remains at rest. Biological mutation does not disrupt the situation as the equilibrium behavior and norm do not change.

Alternatively, biological mutation may alter the cooperation share, $\psi^*(1,\hat{\lambda}) \neq \psi^*(1,\lambda^d)$. Suppose, for example, that the biological mutants strictly prefer to defect at the perfect social norm $\phi^* = 1$ and cooperation costs $\Delta m(\psi^*(1,\lambda))$, resulting in a decrease of the cooperation share, $\psi^*(1,\hat{\lambda}) < \psi^*(1,\lambda)$. Since at the pre-mutation preference distribution λ all individuals maximized biological fitness, cooperating yielded (weakly) greater biological fitness than defecting, $b(1,1,\psi^*(1,\lambda),1) \geq b(0,1,\psi^*(1,\lambda),1)$. The decrease in the cooperation share reduces the costs of cooperation, $\Delta m(\psi^*(1,\hat{\lambda})) < \Delta m(\psi^*(1,\lambda))$, while the social norm and, thus, social disapproval when defecting remain unchanged. It follows that cooperating yields strictly greater biological fitness than defecting after the biological mutation, $b(1,1,\psi^*(1,\hat{\lambda}),1) > b(0,1,\psi^*(1,\hat{\lambda}),1)$. The defecting biological mutants obtain less than average biological fitness, erode, and preferences return to λ. Throughout, the social norm $\phi^* = 1$ does not change. $(\lambda, 1)$ is indeed stable.

Note that if an asymptotically stable perfect-social-norm CE exists at preference distributions λ^d, then λ^d itself constitutes a stable BE for the perfect-social-norm CE. In that case, Proposition 11 classifies which other preference distributions λ constitute stable BE. Consequently, we can assess how different variables and functions relate to the existence of a stable BE and perfect-social-norm CE by identifying when the perfect-social-norm CE of Proposition 4 exists at preference distribution λ^d. Since the dominant preference type θ^d satisfies the specifications of θ in Proposition 6 of Section 4.2.2, the corresponding results apply to λ^d. We expand on these insights by examining small changes that affect cooperation incentives through preferences θ^d. In particular, we investigate small changes in the weight of social approval on biological fitness ρ and social disapproval for hypocrisy h. As a starting point, we suppose that cooperation is incomplete, $\psi^*(1,\lambda^d) \in (0,1)$; otherwise, increases in ρ and h leave the situation unaltered (hypocrisy does not occur, and larger cooperation incentives cannot yield a greater cooperation level).

Proposition 12 *(Comparative results for stable BE and perfect-social-norm CE). Suppose $\psi^*(1,\lambda^d) \in (0,1)$.*

1. *A small increase (decrease) in the weight of social approval on biological fitness ρ favors the existence of a perfect-social-norm CE of Proposition 4 at λ^d if $\frac{\mathrm{d}\Delta m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)} < \frac{(v(1)+h)(\gamma-\rho)}{\psi^*(1,\lambda^d)}$ ($\frac{\mathrm{d}\Delta m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)} > \frac{(v(1)+h)(\gamma-\rho)}{\psi^*(1,\lambda^d)}$).*
2. *A small increase (decrease) in social disapproval for hypocrisy h favors the existence of a perfect-social-norm CE of Proposition 4 at λ^d if $\frac{\mathrm{d}m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)} < \frac{(v(1)+h)(\gamma-\rho)\rho}{\gamma-\psi^*(1,\lambda^d)(\gamma-\rho)}$ ($\frac{\mathrm{d}m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)} > \frac{(v(1)+h)(\gamma-\rho)\rho}{\gamma-\psi^*(1,\lambda^d)(\gamma-\rho)}$).*

Proof. Proposition 20 in Appendix B.3 constitutes the formal equivalent to the above.

Proposition 12 establishes ambiguous effects of small increases in the weight of social approval on biological fitness ρ and social disapproval for hypocrisy h for the existence of an asymptotically stable CE of Proposition 4 at preference distribution λ^d. The described changes increase the cooperation incentives of the dominant preference type θ^d at the perfect social norm $\phi^* = 1$, $\theta^d_s \tilde{v}(1) + \theta^d_s \tilde{h} + \theta^d_p = \rho v(1) + \rho h$, and, consequently, the cooperation share $\psi^*(1,\lambda^d)$. As outlined in Proposition 6 in Section 4.2.2, such an increase in $\psi^*(1,\lambda^d)$ has two countervailing effects: (1) It increases the costs of cooperation, making the perfect social norm less likely an asymptotically stable CE, and (2) it decreases average social disapproval for hypocrisy, $(1-\psi^*)h$, making the perfect social norm more likely an asymptotically stable CE. Depending on the responsiveness of the cooperation costs, $\frac{\mathrm{d}m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)}$, one of the two effects dominates, where less responsiveness generally favors the second. Note that an increase in social disapproval for hypocrisy h requires a less responsive cost curve to be beneficial compared to an increase in the weight of social approval on biological fitness ρ. The reason is that an increase in the weight ρ solely affects the cooperation level $\psi^*(1,\lambda^d)$. In contrast, increasing social disapproval for hypocrisy h also directly impacts cultural fitness differences in the norm non-holder's favor.

These results highlight that although a larger cooperation share $\psi^*(1,\lambda^d)$ may be beneficial for a stable BE and perfect-social-norm CE to exist since it reduces social disapproval for hypocrisy $(1-\psi^*(1,\lambda^d))h$, increasing cooperation incentives through behavioral and biological channels may actually have an adverse effect when the cooperation costs $\Delta m(\cdot)$ are relatively responsive to changes in the cooperation share.

Although the above results prove the potential existence of a stable BE and perfect-social-norm CE $(\lambda^*, 1)$, such a point is never asymptotically stable (given Θ is sufficiently large). In particular, whereas the CE is asymptotically stable, the BE is only stable. The

underlying reason is that biological evolution remains at rest after a biological mutation that leaves the social norm and equilibrium behavior unaltered. The absence of asymptotic stability raises the theoretical possibility of repeated random biological mutations driving society to some preference distribution $\hat{\lambda}$ that is no longer stable or for which the CE $\phi^* = 1$ is no longer stable, both of which may threaten the persistence of cooperation. The following proposition establishes that, under some conditions, this does not pose a threat.

Proposition 13 *(Asymptotically stable set of BE and perfect-social-norm CE). If Conditions 1 and 2b of Proposition 4 hold at λ^d, then $\{(1,\lambda) : \psi^*(1,\lambda) = \psi^*(1,\lambda^d)\}$ is an asymptotically stable set.*

Proof. See Appendix B.3.

Suppose the perfect social norm $\phi^* = 1$ is a CE of Proposition 4 at preference distribution λ^d. Proposition 13 states that the combination of (1) all preference distributions λ mimicking λ^d in terms of equilibrium behavior, $\psi^*(1,\lambda) = \psi^*(1,\lambda^d)$, and (2) the perfect social norm $\phi^* = 1$ form an asymptotically stable set if the difference in social disapproval for non-conformity outweighs average social disapproval for hypocrisy, $\Delta k(1) > (1 - \psi^*(1,\lambda^d))h$. The underlying reason is that in this case, the perfect social norm $\phi^* = 1$ is a CE at any preference distribution λ as long as the cooperation share satisfies $\psi^*(1,\lambda) = \psi^*(1,\lambda^d)$, which introduces some robustness of the perfect-social-norm CE with respect to biological mutations. These insights further underscore the stabilizing role of disapproval for non-conformity $k(\cdot)$ and the potentially destabilizing role of hypocrisy disapproval h. In addition, the importance of a sufficiently large cooperation share ψ^* becomes more prominent.

5.2.4. Biological equilibrium and imperfect-social-norm cultural equilibrium

We continue by investigating when a preference distribution λ and an imperfect social norm $\phi^* \in (0,1)$ constitute a stable BE and CE.

Proposition 14 *(Stable BE and imperfect-social-norm CE). Suppose $\phi^* \in (0,1)$ is a CE of Proposition 7 at λ^d. For any $\lambda \in [0,1]^{|\Theta|}$, (λ,ϕ^*) is a stable rest point if $\phi^* \in (0,1)$ is a CE of Proposition 7 at λ.*

Proof. See Appendix B.3.

We begin by discussing the underlying intuition behind the above proposition. If the social norm ϕ^* is a CE of Proposition 7 at preference distribution λ^d, all norm non-holders strictly prefer to defect and all norm holders to cooperate at social norm ϕ^* and costs of cooperation $\Delta m(\phi^*)$, $\theta_s^d \tilde{v}(\phi^*) < \Delta m(\phi^*) < \theta_s^d \tilde{v}(\phi^*) + \theta_s^d \tilde{h} + \theta_p^d$. Since an individual of preference type θ^d always maximizes biological fitness, cooperating maximizes biological fitness if and only if an individual holds the cooperation norm, $b(1-n,n,\phi^*,\phi^*) < b(n,n,\phi^*,\phi^*)\,\forall n \in \{0,1\}$. If the social norm ϕ^* is a CE of Proposition 7 at preference distribution λ, all individuals strictly prefer to behave accordingly, $\theta_s \tilde{v}(\phi^*) < \Delta m(\phi^*) < \theta_s \tilde{v}(\phi^*) + \theta_s \tilde{h} + \theta_p \Rightarrow (\sigma_1^*,\sigma_0^*) = (1,0)$. Consequently, preference dynamics are at rest. It follows that λ constitutes a BE for which ϕ^* is an asymptotically stable CE.

Suppose some biological mutation to preference distribution $\hat{\lambda}$ occurs. If equilibrium behavior does not change at the post-mutation preference distribution $\hat{\lambda}$, $(\sigma_1^*,\sigma_0^*) = (1,0)$, then the social norm ϕ^* remains a CE and the dynamic system at rest. Consequently, the biological mutation does not disrupt the situation. Alternatively, suppose, for example, that the biological mutants strictly prefer to behave differently. As a result, the social norm ϕ^* may no longer constitute a CE at preference distribution $\hat{\lambda}$. Since the post-mutation preference distribution $\hat{\lambda}$ is close to λ, however, the CE $\hat{\phi}^*$ that society reaches at preference distribution $\hat{\lambda}$ remains close to ϕ^* (recall Lemma 8 in Section 4.2.3). Moreover, since the post-mutation preference distribution $\hat{\lambda}$ and social norm $\hat{\phi}^*$ are close to the pre-mutation values λ and ϕ^*, the post-mutation cooperation share $\psi^*(\hat{\phi},\hat{\lambda})$ and, consequently, the costs of cooperation $\Delta m(\psi^*(\hat{\phi},\hat{\lambda}))$ also remain close to the pre-mutation values $\psi^*(\phi^*,\lambda) = \phi^*$ and $\Delta m(\psi^*(\phi^*,\lambda))$. Hence, after the biological mutation, cooperating still maximizes an individual's biological fitness if and only if the individual holds the cooperation norm, $b(1-n,n,\psi^*(\hat{\phi},\hat{\lambda}),\hat{\phi}) < b(n,n,\psi^*(\hat{\phi},\hat{\lambda}),\hat{\phi})\ \forall n \in \{0,1\}$. 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 cooperation costs do not alter their optimal behavior, implying they still maximize their biological fitness. The biological mutants, however, behave differently. They deviate from biological fitness-maximizing behavior and erode. Preferences return to λ. Since the post-mutation social norm $\hat{\phi}^*$ is close to ϕ^* and the social norm ϕ^* is an asymptotically stable CE at preference distribution λ, cultural evolution reinstates the social norm ϕ^* once preferences return to λ.

Similar to the results of Section 5.2.3, Proposition 14 implies that if an imperfect-social-norm CE exists at preference distribution λ^d, then (λ^d,ϕ^*) forms a stable BE and imperfect-social-norm CE. The proposition then classifies to which other preference distributions λ this also applies. To establish that there may exist a stable BE and imperfect-social-norm CE, it thus remains for us to argue when an imperfect-social-norm CE ϕ^* may exist at preference distribution λ^d. For this purpose, recall that the set of potential imperfect-social-norm CE $I_p(\lambda^d)$ is non-empty if and only if cooperation incentives of norm holders exceed those of norm non-holders, $\theta_s^d \tilde{h} + \theta_p^d > 0$. Substituting for the dominant preference type θ^d yields that the set of potential imperfect-social-norm CE $I_p(\lambda^d)$ is non-empty if and only if social disapproval for hypocrisy exists and affects biological fitness, $\rho h > 0$. Hence, the presence of social disapproval for hypocrisy, $h > 0$, becomes a necessary condition for the existence of a stable BE and imperfect-social-norm CE of Proposition 14. The underlying reason is that social disapproval for hypocrisy is responsible for the wedge in cooperation incentives at the behavioral and biological level, $\theta_s^d \tilde{h} + \theta_p^d = \rho h$.

Proposition 15 *(Potential existence of a stable BE and imperfect-social-norm CE). There exist* $\phi^* \in (0,1)$, $\lambda \in [0,1]^{|\Theta|}$, *and* $\Delta m(\cdot)$ *s.t.* (λ, ϕ^*) *is a stable rest point of Proposition 14 if and only if* $h > 0$.

Proof. See Appendix B.3.

The proposition establishes the potential existence of a stable BE and imperfect-social-norm CE and underscores the criticality of social disapproval for hypocrisy. For each social norm $\phi \in I_p(\lambda^d)$, any cost curve $\Delta m(\cdot)$ inducing Conditions 2 and 3 of Proposition 7 to hold at this social norm and preference distribution λ^d enables the existence of a stable (λ, ϕ^*).

Note that we can infer how (small) changes in different variables and functions may impact an existing stable BE and imperfect-social-norm CE of Proposition 14 by applying the results of Proposition 9. In line with the discussion succeeding Proposition 9, small changes in behavioral incentives do not impact the situation.

By similar reasoning as in Section 5.2.3, any stable (λ^*, ϕ^*) of Proposition 14 is never asymptotically stable, which enables repeatedly occurring biological mutations to threaten the persistence of an imperfect-social-norm CE and, thus, cooperation altogether. Contrary to the perfect-social-norm CE, however, we cannot show that an asymptotically stable set exists that secures the persistence of the imperfect-social-norm CE. This may have severe consequences for society as the imperfect-social-norm CE could eventually vanish by chance. We can show that for this to occur, a biological mutation must lead to the appearance of individuals that are indifferent between both actions at social norm ϕ^* and cooperation costs $\Delta m(\phi^*)$. The following lemma captures this insight formally.

Lemma 3 *(Asymptotically stable set of BE and imperfect-social-norm CE). Consider any CE* ϕ^* *of Proposition 7 at* λ^d. $u(1,n,\phi^*,\phi^*,\theta) \neq u(0,n,\phi^*,\phi^*,\theta)\ \forall \theta \in \Theta, n \in \{0,1\} \Rightarrow \{(\phi^*,\lambda) : (\sigma_1,\sigma_0) = (1,0)\ \forall \sigma \in \Sigma^*(\phi^*,\lambda)\}$ *is an asymptotically stable set.*

Proof. See Appendix B.3.

Lemma 3 states that an asymptotically stable set ensuring the persistence of the imperfect-social-norm CE exists if there could be no preference type θ that induces behavioral indifference, $u(1,n,\phi^*,\phi^*,\theta) \neq u(0,n,\phi^*,\phi^*,\theta)\ \forall n \in \{0,1\}, \theta \in \Theta$. Although this assumption is not consistent with the conceptual approach of this model (since we assume that the set of preference types Θ is arbitrarily large and represents all possible preference types $\mathbb{R}^2_{\geq}$), the lemma illustrates how high the proportion of biological mutations that could threaten the persistence of the imperfect-social-norm CE is.

6. Discussion

The evolutionary analysis highlights the role different social disapproval mechanisms play in preserving and fostering norm-driven cooperation. At the behavioral level, social disapproval for social norm violation and hypocrisy introduce incentives to comply with the social norm and one's personal norm, respectively. Hence, both foster cooperative behavior.

At the cultural level, social disapproval for social norm violation favors the spread of the social norm. Since the share of cooperators among norm holders is generally larger than among norm non-holders, the norm holders are subject to less social disapproval for social norm violation on average, which impacts cultural fitness differences in their favor. Social disapproval for non-conformity supports the persistence of relatively strong social norms even if other forces favor a weakening. At the same time, it potentially traps a relatively weak social norm despite other forces supporting a further spread. By a similar mechanism, it stabilizes a social norm if it is either perfect or absent. Social disapproval for hypocrisy generally hinders the spread of the norm. Since only the carriers of the cooperation norm can experience social disapproval for hypocrisy, they have an evolutionary disadvantage. This insight contrasts with the impact on equilibrium behavior, where greater social concerns for hypocrisy favor cooperation.

At the biological level, preferences for social approval persist due to social disapproval for social norm violation and hypocrisy affecting biological fitness. The preference for self-approval compensates for the inability to overlook the full extent of actions regarding social disapproval for hypocrisy. If individuals could oversee the full extent, $h = (1+\delta)\tilde{h}$, then no personal concerns are necessary for the dominant preference type to maximize biological fitness. Individuals who hold the cooperation norm would follow their personal norms only to avoid social disapproval for 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 for social norm violation. Cultural evolution is solely driven by conformity, implying society could only reach a stable perfect- or no-social-norm CE. The potential heterogeneity in personal norms vanishes. If social disapproval for non-conformity were also absent, cultural evolution would be subject to random walk. For deterministic cultural evolution to occur, society requires some other mechanism such as institutional pressure (see, e.g., Gintis, 2003b; Mengel, 2008) or conformity bias in social learning (see, e.g., Henrich and Boyd, 1998, 2001; Chudek and Henrich, 2011; Nordblom and Žamac, 2012; Michaeli and Spiro, 2015).

These insights underscore the importance of one of the primary motivations for developing this paper's dynamic model: the mutual endogenization of norms and preferences to draw a more complete picture of the underlying dynamic system. When preferences and norms are endogenous, social disapproval for social norm violation alone does not determine norms, whereas if preferences are exogenous, it does.

Our results provide some rationale for the existence of heterogeneous preferences leading to pro-social behavior, an observation generally in line with empirical findings (see, e.g., Fisman et al., 2007). In particular, our analysis indicates that individuals can differ vastly regarding their biological characteristics as long as all behave optimally regarding biological fitness given the aggregate

outcome. Within the scope of our model, this holds if preferences are distributed such that the resulting equilibrium behavior at some social norm is as if society was homogeneous with preferences that maximize biological fitness. These results also contribute to the literature on gene-culture co-evolution (see, among others, Gintis, 2003a, 2011; Richerson and Boyd, 2010; Richerson et al., 2010; Chudek and Henrich, 2011), as they highlight how the existing culture (social norm) shapes the genes (preferences) that may prevail in equilibrium.

Throughout, the analysis concerns the asymptotic stability of CE, but only stability (in the sense of Lyapunov) of BE. The underlying reason is that although biological mutation may alter preferences, such alterations may not disrupt the prevailing social norm or equilibrium behavior. The biological mutation has no effect in that case, and the dynamic system remains at rest. However, if it does have an effect, our analysis indicates that the biological mutants erode and society returns to the original social norm and equilibrium behavior. Nevertheless, the absence of asymptotic stability raises the theoretical possibility of repeated random biological mutations driving preferences to some distribution that is no longer stable. We have shown that, under some conditions, this does not threaten the persistence of a perfect-social-norm CE, implying our model can explain the endurance of (possibly incomplete) norm-driven cooperation. However, we cannot infer the same for an imperfect-social-norm CE, suggesting it could vanish by chance. Nevertheless, our results indicate that the probability of this occurring at any point in time seems rather small, since, of all preference types that can appear, the biological mutation must give rise to precisely those preference types that induce indifference between both actions. This reasoning seems especially appropriate when interpreting the model of this paper as a continuous approximation of a discrete system with finitely many individuals as well as an infinite set of possible preference types ($\Theta = \mathbb{R}^2_{\geq 0}$). Moreover, accrediting a small probability to the biological mutations that may trigger the erosion of the imperfect social norm can explain why imperfect social norms have existed so far, even if they cannot be ensured to prevail.

7. 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 existence of different stable equilibria varying regarding their social norms and cooperation levels. Although in equilibrium, preferences are potentially heterogeneous, behavior is as if they were homogeneous. Social disapproval for social norm violation incentivizes individuals 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. Social disapproval for hypocrisy introduces cooperation incentives at the behavioral and biological levels. Thereby, it enables the existence of stable long-run equilibria characterized by heterogeneous personal norms and behavior. However, it negatively impacts the cultural fitness of individuals who defect despite holding cooperation-prescribing personal norms, possibly hindering the preservation of a perfect social norm if cooperation is very costly.

The model of this paper builds on some notable assumptions that need further investigation. We assume that the reproduction of norms and preferences depends 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 perspective is in line with traditional approaches from *evolutionary game theory* that consider material payoff as the sole determinant of reproductive fitness. The underlying idea in this paper more closely aligns with approaches from *cultural evolutionary models*. The assumption is 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 an endogenous process to society. Future research in that direction needs to complement this paper.

Moreover, we made some assumptions regarding the relationship between cultural and biological evolution. Particularly, we assumed that social approval has a more significant impact on cultural than biological fitness. In addition, we assumed that norms are equally distributed across all preference types. In reality, it seems plausible that the share of norm holders differs across preference types, possibly because either individuals may be more inclined to adopt norms from others that are similar to them or certain preference types are more likely to adopt pro-social norms. In the latter case, it is an open question of how preferences condition the adoption of norms: Is it rather the individuals who likely follow norms who adopt them? Or is it precisely these individuals who do not, since they then may cooperate even when it is sub-optimal from a fitness perspective? Further research needs to address these questions, as well as how our results change when weakening the assumptions.

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 in other situations. Since gossip does not create any incentive problems regarding one's own optimal behavior, the results of this paper can 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. In Section 2, we argued that a positive probability of being detected when lying and severe material or social costs when being detected may cause truth-telling to be the best response. Nevertheless, further work incorporating communication as a behavioral dimension needs to complement this paper.[15]

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

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The work was developed as part of the *ZumWert* project. I want to thank the University of Kassel for pursuing its profile building program, which made this project possible. Moreover, I thank Georg von Wangenheim and two anonymous referees for their valuable and detailed comments and suggestions. I also thank Marlene Batzke and Andreas Ernst for insightful discussions.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.geb.2024.07.006.

Appendix B. Proofs and additional formal results

B.1. Behavior

Lemma 4. *For all $\lambda \in [0,1]^{|\Theta|}$ and $\phi \in [0,1]$, the set of all NE is convex and asymptotically stable under a best-response dynamics.*

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

Lemma 5. *For all $\lambda \in [0,1]^{|\Theta|}$, $\phi \in [0,1]$, $n \in \{0,1\}$, and $\theta \in \text{supp}(\lambda)$, $\sigma \in \Sigma^*(\phi,\lambda)$ implies:*

1. *$\sigma_{n,\theta} = 1$ if $n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi) > \Delta m(\phi\sigma_1 + (1-\phi)\sigma_0)$ and*
2. *$\sigma_{n,\theta} = 0$ if $n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi) < \Delta m(\phi\sigma_1 + (1-\phi)\sigma_0)$.*

Proof. Consider any $\sigma \in \Sigma^*(\phi,\lambda)$. Consider the difference in utilities from cooperation and defection for individuals with personal norm n and preference type θ: $\Delta u_{n,\theta} = u(1,n,\psi,\phi,\theta) - u(0,n,\psi,\phi,\theta) = n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi) - \Delta m(\phi\sigma_1 + (1-\phi)\sigma_0)$. Then $n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi) > \Delta m(\phi\sigma_1 + (1-\phi)\sigma_0)$ implies that cooperation is strictly preferred to defection. Thus, $\sigma_{n,\theta} = 1$ must be true in any NE. Analogously, $n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi) < \Delta m(\phi\sigma_1 + (1-\phi)\sigma_0)$ implies that defecting is strictly preferred. Thus, $\sigma_{n,\theta} = 0$ must be true in any NE. □

Lemma 6. *For all $\lambda \in [0,1]^{|\Theta|}$, $\phi \in [0,1]$, and $\hat{\sigma}, \check{\sigma} \in \Sigma^*(\phi,\lambda)$, $\phi\hat{\sigma}_1 + (1-\phi)\hat{\sigma}_0 = \phi\check{\sigma}_1 + (1-\phi)\check{\sigma}_0$.*

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

Lemma 7. *For all $\lambda \in [0,1]^{|\Theta|}$, $\check{\theta}, \hat{\theta} \in \text{supp}(\lambda)$, $\phi \in [0,1]$, and $\check{n}, \hat{n} \in \{0,1\}$, $(\check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\phi) > \hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\phi)) \Rightarrow (\sigma_{\check{n},\check{\theta}} \geq \sigma_{\hat{n},\hat{\theta}} \; \forall \sigma \in \Sigma^*(\phi,\lambda))$.*

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

Lemma 8. *For all $\lambda \in [0,1]^{|\Theta|}$, $\psi^*(\phi,\lambda)$ is non-decreasing in ϕ.*

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

Lemma 9. *For all $\lambda \in [0,1]^{|\Theta|}$, $\phi \in [0,1]$, and $\sigma \in \Sigma^*(\phi,\lambda)$, $\sigma_0 \leq \sigma_1$.*

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

Lemma 10. *For all $\lambda \in [0,1]^{|\Theta|}$ and $\phi \in [0,1]$, $(\hat{\theta}_p + \hat{\theta}_s \tilde{h} + \hat{\theta}_s \tilde{v}(\phi) \neq \tilde{\theta}_s \tilde{v}(\phi)\ \forall \hat{\theta}, \tilde{\theta} \in \text{supp}(\lambda)) \Rightarrow ((\hat{\sigma}_1, \hat{\sigma}_0) = (\check{\sigma}_1, \check{\sigma}_0)\ \forall \hat{\sigma}, \check{\sigma} \in \Sigma^*(\phi,\lambda))$.*

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

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

Lemma 11. *For all $\lambda \in [0,1]^{|\Theta|}$ and $\phi \in [0,1]$, (for all $\hat{\theta}, \check{\theta} \in \text{supp}(\lambda)$ and $\hat{n}, \check{n} \in \{0,1\}$, $\hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\phi) = \check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\phi) \Rightarrow \hat{\theta} = \check{\theta}$ and $\hat{n} = \check{n}) \Rightarrow (\Sigma^*(\phi,\lambda)$ is a singleton).*

Proof. Consider any $\lambda \in [0,1]^{|\Theta|}$ and $\phi \in [0,1]$. If for all $\hat{\theta}, \check{\theta} \in \text{supp}(\lambda)$ and $\hat{n}, \check{n} \in \{0,1\}$, $\hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\phi) = \check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\phi) \Rightarrow \hat{\theta} = \check{\theta} \wedge \hat{n} = \check{n}$, then there is at most one pairing of $\bar{n} \in \{0,1\}$ and $\bar{\theta} \in \text{supp}(\lambda)$ s.t. $\bar{n}(\bar{\theta}_p + \bar{\theta}_s \tilde{h}) + \bar{\theta}_s \tilde{v}(\phi) = \Delta m(\psi^*(\phi,\lambda))$.

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

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

Lemma 12. *For all λ and $\phi \in [0,1]$, there is a neighborhood U of ϕ s.t. $\forall \hat{\phi} \in U \setminus \{\phi\}$, $\Sigma^*(\hat{\phi},\lambda)$ is a singleton.*

Proof. Consider any $\phi \in [0,1]$. Let $\epsilon > 0$ be s.t. $\forall \hat{\theta}, \check{\theta} \in \Theta$ and $\hat{n}, \check{n} \in \{0,1\}$, $\hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\phi) > \check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\phi) \Rightarrow \hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\hat{\phi}) > \check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\hat{\phi})$ for all $\hat{\phi} \in (\phi - \epsilon, \phi + \epsilon)$. Such ϵ exists due to continuity of $\tilde{v}$. Moreover, consider any $\hat{n}, \check{n} \in \{0,1\}$ and $\hat{\theta}, \check{\theta} \in \Theta$ s.t. $\hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\phi) = \check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\phi)$. ($\check{\theta} \neq \hat{\theta}$ or $\check{n} \neq \hat{n} \Rightarrow \hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\hat{\phi}) \neq \check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\hat{\phi})\ \forall \hat{\phi} \neq \phi$. Hence, for all $\lambda \in [0,1]^{|\Theta|}$ and $\hat{\phi} \in (\phi - \epsilon, \phi + \epsilon) \setminus \{\phi\}$, (for all $\hat{\theta}, \check{\theta} \in \text{supp}(\lambda)$ and $\hat{n}, \check{n} \in \{0,1\}$, $\hat{n}(\hat{\theta}_p + \hat{\theta}_s \tilde{h}) + \hat{\theta}_s \tilde{v}(\phi) = \check{n}(\check{\theta}_p + \check{\theta}_s \tilde{h}) + \check{\theta}_s \tilde{v}(\phi) \Rightarrow \hat{\theta} = \check{\theta}$ and $\hat{n} = \check{n}$). Lemma 11 implies that for all $\lambda \in [0,1]^{|\Theta|}$ and $\hat{\phi} \in (\phi - \epsilon, \phi + \epsilon) \setminus \{\phi\}$, $\Sigma^*(\hat{\phi},\lambda)$ is a singleton. □

Lemma 13. *For all $\lambda \in [0,1]^{|\Theta|}$, $\lim_{x \to 1}(\psi^*(x,\lambda)) = \psi^*(1,\lambda)$.*

Proof. Lemma 8 proves that ψ^* is non-decreasing in ϕ. Thus, $\psi^*(\phi,\lambda) \leq \psi^*(1,\lambda)$ for all $\phi < 1$. It remains to be shown that $\forall \epsilon > 0$, $\exists \xi > 0$ s.t. $(\phi > 1 - \xi \Rightarrow \psi^*(\phi,\lambda) > \psi^*(\phi,\lambda) - \epsilon)$.

Note that $\sum_{\theta \in \text{supp}(\lambda) \text{ s.t. } \theta_p + \theta_s \tilde{s} + \theta_s \tilde{v}(1) \geq \Delta m(\psi^*(1,\lambda))} \lambda_\theta = \psi^*(1,\lambda)$ is true. Moreover, it implies that $\sum_{\theta \in \text{supp}(\lambda) \text{ s.t. } \theta_p + \theta_s \tilde{s} + \theta_s \tilde{v}(1) > \Delta m(\psi^*(1,\lambda) - \epsilon)} \lambda_\theta \geq \psi^*(1,\lambda) > \psi^*(1,\lambda) - \epsilon$. Let $\xi > 0$ be s.t. for all $\phi > 1 - \xi$, $\phi \sum_{\theta \in \text{supp}(\lambda) \text{ s.t. } \theta_p + \theta_s \tilde{s} + \theta_s \tilde{v}(\phi) > \Delta m(\psi^*(1,\lambda) - \epsilon)} \lambda_\theta > \psi^*(1,\lambda) - \epsilon$. Such a ξ exists due to continuity of $\tilde{v}$.

At any $\phi > 1 - \xi$, it must be true that $\psi^*(\phi,\lambda) > \psi^*(1,\lambda) - \epsilon$. Assume by contradiction that $\psi^*(\phi,\lambda) \leq \psi^*(1,\lambda) - \epsilon$. $\psi^*(\phi,\lambda) \leq \psi^*(1,\lambda) - \epsilon \Rightarrow \Delta m(\psi^*(\phi,\lambda)) \leq \Delta m(\psi^*(1,\lambda) - \epsilon) \Rightarrow (\sum_{\theta \in \text{supp}(\lambda) \text{ s.t. } \theta_p + \theta_s \tilde{s} + \theta_s \tilde{v}(\phi) > \Delta m(\psi^*(1,\lambda) - \epsilon)} \lambda_\theta > \frac{\psi^*(1,\lambda) - \epsilon}{\phi} \Rightarrow \sum_{\theta \in \text{supp}(\lambda) \text{ s.t. } \theta_p + \theta_s \tilde{s} + \theta_s \tilde{v}(\phi) > \Delta m(\psi^*(\phi,\lambda))} \lambda_\theta > \frac{\psi^*(1,\lambda) - \epsilon}{\phi}) \Rightarrow \sigma_1^* > \frac{\psi^*(1,\lambda) - \epsilon}{\phi}$ for all $\sigma^* \in \Sigma^*(\phi,\lambda) \Rightarrow \psi^*(\phi,\lambda) \geq \sigma_1^* \phi > \psi^*(1,\lambda) - \epsilon$. We reached a contradiction. Therefore $\psi^*(\phi,\lambda) > \psi^*(1,\lambda) - \epsilon$ for any $\phi > 1 - \xi$. Lemma 13 is true. □

Lemma 14. *For all $\lambda \in [0,1]^{|\Theta|}$, $\lim_{x \to 1}(\min_{\sigma \in \Sigma^*(x,\lambda)}(\sigma_1)) = \lim_{x \to 1}(\max_{\sigma \in \Sigma^*(x,\lambda)}(\sigma_1)) = \psi^*(1,\lambda)$.*

Proof. From $\sigma_1 > \sigma_0$ for all $\sigma \in \Sigma^*(\phi,\lambda)$ and $\psi = \phi\sigma_1 + (1-\phi)\sigma_0$ follows that for all $\phi \in [0,1)$ and $\sigma \in \Sigma^*(\phi,\lambda)$: $\min_{\sigma \in \Sigma^*(\phi,\lambda)}(\sigma_1) \geq \psi^*(\phi,\lambda)$ and $\max_{\sigma \in \Sigma^*(x,\lambda)}(\sigma_1) \leq \frac{\psi^*(\phi,\lambda)}{\phi}$. Proposition 14 follows from $\lim_{x \to 1}(\psi^*(x,\lambda)) = \lim_{x \to 1}(\frac{\psi^*(x,\lambda)}{x}) = \psi^*(1,\lambda)$. □

Lemma 15. *Consider any* $\lambda \in [0,1]^{|\Theta|}$ *and* $\phi \in [0,1]$. $\forall \epsilon > 0\ \exists \xi > 0$ *s.t.* $|\phi - \hat{\phi}| < \xi \Rightarrow |\psi^*(\phi,\lambda) - \psi^*(\hat{\phi},\lambda)| < \epsilon$.

Proof. Consider any $\lambda \in [0,1]^{|\Theta|}$, $\phi \in [0,1]$, $\sigma \in \Sigma^*(\phi,\lambda)$. Suppose by contradiction that there is $\epsilon > 0$ s.t. for all $\xi > 0$, $|\phi - \hat{\phi}| < \xi$ and $|\psi^*(\phi,\lambda) - \psi^*(\hat{\phi},\lambda)| \geq \epsilon$ for some $\hat{\phi}$. Consider any such $\epsilon > 0$.

$\psi^*(\phi,\lambda) = \psi^*(\hat{\phi},\lambda)$ clearly leads to a contradiction. Next, suppose $\psi^*(\hat{\phi},\lambda) \leq \psi^*(\phi,\lambda) - \epsilon$ (analogously for $\psi^*(\hat{\phi},\lambda) \geq \psi^*(\phi,\lambda) + \epsilon$). Since ψ^* is non-decreasing in ϕ (recall Lemma 8), this can only be true if $\hat{\phi} < \phi$. Consider any $\theta \in \text{supp}(\lambda), n \in \{0,1\}$ s.t. $\theta_s \tilde{v}(\phi) + (\theta_s \tilde{h} + \theta_p)n \geq \Delta m(\psi^*(\phi,\lambda)) > \Delta m(\psi^*(\phi,\lambda) - \epsilon)$. Let ξ be so small that $\theta_s \tilde{v}(\hat{\phi}) + (\theta_s \tilde{h} + \theta_p)n > \Delta m(\psi^*(\phi,\lambda) - \epsilon)$ for all $\hat{\phi} \in (\phi - \xi, \phi)$. Since (a) $\psi^*(\hat{\phi},\lambda) < \psi^*(\phi,\lambda) - \epsilon$ and (b) Δm increasing in ψ, $\theta_s \tilde{v}(\hat{\phi}) + (\theta_s \tilde{h} + \theta_p)n > \Delta m(\psi^*(\hat{\phi},\lambda))$ for all $\hat{\phi} \in (\phi - \xi, \phi)$. For all $\theta \in \text{supp}(\lambda), n \in \{0,1\}, \hat{\phi} \in (\phi, \phi - \xi), \sigma \in \Sigma^*(\phi,\lambda), \hat{\sigma} \in \Sigma^*(\hat{\phi},\lambda)$, $\sigma_{n,\theta} > 0 \Rightarrow \theta_s \tilde{v}(\phi) + (\theta_s \tilde{h} + \theta_p)n \geq \Delta m(\psi^*(\phi,\lambda)) \Rightarrow \theta_s \tilde{v}(\hat{\phi}) + (\theta_s \tilde{h} + \theta_p)n > \Delta m(\psi^*(\hat{\phi},\lambda)) \Rightarrow \hat{\sigma}_{n,\theta} = 1 \Rightarrow \sigma_n \leq \hat{\sigma}_n$. Consequently, $\psi^*(\phi,\lambda) - \psi^*(\hat{\phi},\lambda) = \sigma_{1,\theta}\phi + \sigma_{0,\theta}(1-\phi) - \hat{\sigma}_{1,\theta}\hat{\phi} - \hat{\sigma}_{0,\theta}(1-\hat{\phi}) \leq \sigma_{1,\theta}\phi + \sigma_{0,\theta}(1-\phi) - \sigma_{1,\theta}\hat{\phi} - \sigma_{0,\theta}(1-\hat{\phi}) = \sigma_{1,\theta}(\phi - \hat{\phi}) + \sigma_{0,\theta}(\hat{\phi} - \phi) < \sigma_1 \xi$. Hence, for $\xi \leq \frac{\epsilon}{\sigma_1}$, $\phi - \hat{\phi} < \xi \Rightarrow \psi^*(\phi,\lambda) - \psi^*(\hat{\phi},\lambda) < \epsilon$. We have reached a contradiction, implying that the lemma is true. □

Lemma 16. *Consider any* $\lambda \in [0,1]^{|\Theta|}$ *and* $\phi \in [0,1]$. $\forall \epsilon > 0\ \exists \xi > 0$ *s.t.* $\forall \hat{\lambda} \in [0,1]^{|\Theta|}$: $\sum_{\theta \in \Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \xi \Rightarrow |\psi^*(\phi,\lambda) - \psi^*(\phi,\hat{\lambda})| < \epsilon$.

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

Suppose that $\psi^*(\phi,\lambda) > \psi^*(\phi,\hat{\lambda})$ (analogously for $\psi^*(\phi,\lambda) < \psi^*(\phi,\hat{\lambda})$). For all $\theta \in \text{supp}(\lambda)$ and $n \in \{0,1\}$, $\psi^*(\phi,\lambda) > \psi^*(\phi,\hat{\lambda}) \Rightarrow \Delta m(\psi^*(\phi,\lambda)) > \Delta m(\psi^*(\phi,\hat{\lambda})) \Rightarrow (n(\theta_s \tilde{h} + \theta_p) + \theta_s \tilde{v}(\phi) \geq \Delta m(\psi^*(\phi,\lambda)) \Rightarrow n(\theta_s \tilde{h} + \theta_p) + \theta_s \tilde{v}(\phi) > \Delta m(\psi^*(\phi,\hat{\lambda}))) \Rightarrow \hat{\sigma}_{n,\theta} \geq \sigma_{n,\theta} \Rightarrow \psi^*(\phi,\hat{\lambda}) \geq \phi \sum_{\theta \in \text{supp}(\lambda)} \hat{\lambda}_\theta \sigma_{1,\theta} + (1-\phi) \sum_{\theta \in \text{supp}(\lambda)} \hat{\lambda}_\theta \sigma_{0,\theta}$. Moreover, $\psi^*(\phi,\lambda) = \phi \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \sigma_{1,\theta} + (1-\phi) \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \sigma_{0,\theta}$. $0 < \psi^*(\phi,\lambda) - \psi^*(\phi,\hat{\lambda}) \leq \phi \sum_{\theta \in \text{supp}(\lambda)} (\lambda_\theta - \hat{\lambda}_\theta)\sigma_{1,\theta} + (1-\phi) \sum_{\theta \in \text{supp}(\lambda)} (\lambda_\theta - \hat{\lambda}_\theta)\sigma_{0,\theta} \leq \sum_{\theta \in \text{supp}(\lambda)} (\lambda_\theta - \hat{\lambda}_\theta) \leq \sum_{\theta \in \Theta} (\lambda_\theta - \hat{\lambda}_\theta) < \xi < \epsilon$. Hence, $\sum_{\theta \in \Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \xi$ implies that $|\psi^*(\phi,\lambda) - \psi^*(\phi,\hat{\lambda})| < \xi$. Thus, Lemma 16 is true. □

Lemma 17. *Consider any* $\phi \in (0,1)$ *and* $\lambda \in [0,1]^{|\Theta|}$ *s.t.* $\forall \theta, \bar{\theta} \in \text{supp}(\lambda)$, $\bar{\theta}_s \tilde{v}(\phi) \neq \theta_s \tilde{v}(\phi) + \theta_s \tilde{h} + \theta_p$. $\forall \epsilon > 0\ \exists \xi > 0$ *s.t.* $\forall \hat{\lambda} \in [0,1]^{|\Theta|}$, $\sum_{\theta \in \Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \xi$ *implies that for all* $\sigma \in \Sigma^*(\phi,\lambda)$ *and* $\hat{\sigma} \in \Sigma^*(\phi,\hat{\lambda})$, $|\sigma_n - \hat{\sigma}_n| < \epsilon\ \forall n \in \{0,1\}$.

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

First, we look at $\psi^*(\phi,\lambda) = \psi^*(\phi,\hat{\lambda})$. Let $n \in \{0,1\}$ be s.t. for all $\theta \in \text{supp}(\lambda)$, $n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi) \neq \Delta m(\psi^*(\phi,\lambda))$. Such an n exists since $\bar{\theta}_s \tilde{v}(\phi) \neq \check{\theta}_s \tilde{v}(\phi) + \check{\theta}_s \tilde{h} + \check{\theta}_p$ for all $\check{\theta}, \bar{\theta} \in \text{supp}(\lambda)$. For all $\theta \in \text{supp}(\lambda)$, $((n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi) > \Delta m(\psi^*(\phi,\lambda)) \Rightarrow n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi) > \Delta m(\psi^*(\phi,\hat{\lambda})))$ and $(n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi) < \Delta m(\psi^*(\phi,\lambda)) \Rightarrow n(\theta_p + \theta_s \tilde{h}) + \theta_s \tilde{v}(\phi) < \Delta m(\psi^*(\phi,\hat{\lambda})))) \Rightarrow \hat{\sigma}_{n,\theta} = \sigma_{n,\theta}$. $(\hat{\sigma}_{n,\theta} = \sigma_{n,\theta} \forall \theta \in \text{supp}(\lambda) \wedge \sigma_n = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \sigma_{n,\theta} \wedge \hat{\sigma}_n = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \hat{\sigma}_{n,\theta} + \sum_{\theta \in \Theta} (\hat{\lambda}_\theta - \lambda_\theta)\hat{\sigma}_{n,\theta}) \Rightarrow \hat{\sigma}_n = \sigma_n + \sum_{\theta \in \Theta} (\hat{\lambda}_\theta - \lambda_\theta)\hat{\sigma}_{n,\theta}$. $\hat{\sigma}_n - \sigma_n = \sum_{\theta \in \Theta} (\hat{\lambda}_\theta - \lambda_\theta)\hat{\sigma}_{n,\theta} \Rightarrow |\sigma_n - \hat{\sigma}_n| = \sum_{\theta \in \Theta} (|\hat{\lambda}_\theta - \lambda_\theta|)\hat{\sigma}_{n,\theta} < \sum_{\theta \in \Theta} (|\hat{\lambda}_\theta - \lambda_\theta|) < \xi < \epsilon \Rightarrow |\sigma_n - \hat{\sigma}_n| < \epsilon$. Moreover, let $x = |n - 1 + \phi|$. Recall $\xi < \epsilon$, $\xi \frac{x}{1-x} < \epsilon$, and $|\sigma_n - \hat{\sigma}_n| < \epsilon$. $\psi^*(\phi,\lambda) = \psi^*(\phi,\hat{\lambda}) \Rightarrow x\sigma_n + (1-x)\sigma_{1-n} = x\hat{\sigma}_n + (1-x)\hat{\sigma}_{1-n} \Rightarrow |\sigma_{1-n} - \hat{\sigma}_{1-n}| = |\sigma_n - \hat{\sigma}_n| \times \frac{x}{1-x} < \xi \times \frac{x}{1-x} = \epsilon \times \min\{\frac{\phi}{1-\phi}, \frac{1-\phi}{\phi}\} \times \frac{x}{1-x}$. $(\min\{\frac{\phi}{1-\phi}, \frac{1-\phi}{\phi}\} < 1$ and $(\frac{x}{1-x} = \min\{\frac{\phi}{1-\phi}, \frac{1-\phi}{\phi}\}$ or $\frac{x}{1-x} = 1/\min\{\frac{\phi}{1-\phi}, \frac{1-\phi}{\phi}\}))) \Rightarrow \frac{x}{1-x} \times \min\{\frac{\phi}{1-\phi}, \frac{1-\phi}{\phi}\} \leq 1 \Rightarrow \xi \frac{x}{1-x} < \epsilon$. Thus, $|\sigma_{1-n} - \hat{\sigma}_{1-n}| < \epsilon$.

Second, we look at $\psi^*(\phi,\lambda) < \psi^*(\phi,\hat{\lambda})$. For all $\theta \in \text{supp}(\lambda)$ and $n \in \{0,1\}$, $\psi^*(\phi,\lambda) < \psi^*(\phi,\hat{\lambda}) \Rightarrow \Delta m(\psi^*(\phi,\lambda)) < \Delta m(\psi^*(\phi,\hat{\lambda})) \Rightarrow (n(\theta_s \tilde{h} + \theta_p) + \theta_s \tilde{v}(\phi) \leq \Delta m(\psi^*(\phi,\lambda)) \Rightarrow n(\theta_s \tilde{h} + \theta_p) + \theta_s \tilde{v}(\phi) < \Delta m(\psi^*(\phi,\hat{\lambda}))) \Rightarrow \hat{\sigma}_{n,\theta} \leq \sigma_{n,\theta} \Rightarrow \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \hat{\sigma}_{n,\theta} \leq \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \sigma_{n,\theta} = \sigma_n$. $\hat{\sigma}_n = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \hat{\sigma}_{n,\theta} + \sum_{\theta \in \Theta} (\hat{\lambda}_\theta - \lambda_\theta)\hat{\sigma}_{n,\theta}$. $(\sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \hat{\sigma}_{n,\theta} \leq \sigma_n \wedge \sum_{\theta \in \Theta} (|\hat{\lambda}_\theta - \lambda_\theta|)\hat{\sigma}_{n,\theta} < \xi) \Rightarrow \hat{\sigma}_n < \sigma_n + \xi$. Thus, $\hat{\sigma}_1 < \sigma_1 + \xi$ and $\hat{\sigma}_0 < \sigma_0 + \xi$. Let $n \in \{0,1\}$ be s.t. $\hat{\sigma}_n > \sigma_n$. Such an n must exist since otherwise $\psi^*(\phi,\lambda) < \psi^*(\phi,\hat{\lambda})$ cannot be true. It follows that $\sigma_n < \hat{\sigma}_n < \sigma_n + \xi \Rightarrow |\sigma_n - \hat{\sigma}_n| < \xi < \epsilon$. Let $x = |n - 1 + \phi|$. $\psi^*(\phi,\lambda) < \psi^*(\phi,\hat{\lambda}) \Rightarrow x\sigma_n + (1-x)\sigma_{1-n} < x\hat{\sigma}_n + (1-x)\hat{\sigma}_{1-n}$. $\hat{\sigma}_n < \sigma_n + \xi \Rightarrow x\sigma_n + (1-x)\sigma_{1-n} < x\sigma_n + x\xi + (1-x)\hat{\sigma}_{1-n} \Rightarrow \sigma_{1-n} - \frac{x}{1-x}\xi < \hat{\sigma}_{1-n} < \sigma_{1-n} + \xi$. By same reasoning as above, $\xi \frac{x}{1-x} < \epsilon$. $(\xi < \epsilon$ and $\xi \frac{x}{1-x} < \epsilon) \Rightarrow |\sigma_n - \hat{\sigma}_n| < \epsilon$. 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 $\epsilon > 0$, $\sum_{\theta \in \Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \epsilon \times \min\{\frac{\phi}{1-\phi}, \frac{1-\phi}{\phi}\}$ implies that $|\sigma_n - \hat{\sigma}_n| < \epsilon$. Thus, the lemma is indeed true. □

Proposition 16. *Consider any specification of* θ, $\tilde{v}(\cdot)$, $\tilde{h}$, $\Delta m(\cdot)$, ϕ, *and* λ *s.t.* $\text{supp}(\lambda) = \{\theta\}$.

1. *For all* $x, y \in \mathbb{R}_{\geq 0}$, $(\theta_p = x \Rightarrow \psi^*(\phi,\lambda) = y) \Rightarrow (\theta_p > x \Rightarrow \psi^*(\phi,\lambda) \geq y)$.
2. *For all* $x, y \in \mathbb{R}_{\geq 0}$, $(\theta_s = x \Rightarrow \psi^*(\phi,\lambda) = y) \Rightarrow (\theta_s > x \Rightarrow \psi^*(\phi,\lambda) \geq y)$.
3. *For all* $x, y \in \mathbb{R}_{\geq 0}$, $(\tilde{h} = x \Rightarrow \psi^*(\phi,\lambda) = y) \Rightarrow (\tilde{h} > x \Rightarrow \psi^*(\phi,\lambda) \geq y)$.
4. *For all* $x(z)$ *and* $y \in \mathbb{R}_{\geq 0}$, $(\tilde{v}(z) = x(z) \Rightarrow \psi^*(\phi,\lambda) = y) \Rightarrow (\tilde{v}(z) > x(z) \Rightarrow \psi^*(\phi,\lambda) \geq y)$.
5. *For all* $x(z)$ *and* $y \in \mathbb{R}_{\geq 0}$, $(\tilde{v}(z) = x(z) \Rightarrow \psi^*(\phi,\lambda) = y) \Rightarrow (\tilde{v}(z) > x(z) \Rightarrow \psi^*(\phi,\lambda) \geq y)$.
6. *For all* $x(z)$ *and* $y \in \mathbb{R}_{\geq 0}$, $(\Delta m(z) = x(z) \Rightarrow \psi^*(\phi,\lambda) = y) \Rightarrow (\Delta m(z) < x(z) \Rightarrow \psi^*(\phi,\lambda) \geq y)$.

Proof. Consider any specification of the model and $\phi \in [0,1]$. Since $\text{supp}(\lambda)$ is a singleton, $\Sigma^*(\phi,\lambda)$ is a singleton (see Lemma 11). We can derive the following conditions for equilibrium behavior:

1. $\sigma_1^* = \sigma_0^* = \psi^* = 0$ if $\theta_s \tilde{v}(\phi) + \theta_s \tilde{h} + \theta_p \leq \Delta m(0)$,
2. $\sigma_0^* = 0, \sigma_1^* \in [0,1], \psi^* \in [0,\phi]$ s.t. $\Delta m(\psi^*) = \theta_s \tilde{v}(\phi) + \theta_s \tilde{h} + \theta_p$ if $\Delta m(0) \leq \theta_s \tilde{v}(\phi) + \theta_s \tilde{h} + \theta_p \leq \Delta m(\phi)$,
3. $\sigma_0^* = 0, \sigma_1^* = 1, \psi^* = \phi$ if $\theta_s \tilde{v}(\phi) \leq \Delta m(\phi) \leq \theta_s \tilde{v}(\phi) + \theta_s \tilde{h} + \theta_p$,
4. $\sigma_0^* \in [0,1], \sigma_1^* = 1, \psi^* \in [\phi,1]$ s.t. $\Delta m(\psi^*) = \theta_s \tilde{v}(\phi) + \theta_s \tilde{h} + \theta_p$ if $\Delta m(\phi) \leq \theta_s \tilde{v}(\phi) + \theta_s \tilde{h} + \theta_p \leq \Delta m(1)$, and
5. $\sigma_1^* = \sigma_0^* = \psi^* = 1$ if $\Delta m(1) \leq \theta_s \tilde{v}(\phi) \geq$.

Suppose, for example, $\sigma_0^* = 0, \sigma_1^* \in [0,1], \psi^* \in [0,\phi]$ (case 2). A change in the model to some larger θ_s, θ_p, $\tilde{h}$, $\tilde{v}(\cdot)$, or $-\Delta m(\cdot)$ leads to the if condition of case 2 being either violated or not. In the former, one of the cases 3, 4, or 5 applies, implying $\psi^* \geq \phi$. In the latter case, $\Delta m(\psi^*) = \theta_s \tilde{v}(\phi) + \theta_s \tilde{h} + \theta_p$ holds before and after the change. Since $\Delta m'(x)\ \forall x$, greater θ_s, θ_p, $\tilde{h}$, $\tilde{v}(\cdot)$, or $-\Delta m(\cdot)$ must, thus, be accompanied by an increase in ψ^*. The arguments for the remaining cases are analog. □

B.2. Norms

Proof of Proposition 3. Consider any $\lambda \in [0,1]^{|\Theta|}$. $\phi = 0$ is always a rest point of norm dynamics (Definition 7). It remains to be shown that $\phi = 0$ is asymptotically stable. $\phi = 0$ is asymptotically stable if for all $\hat{\phi}$ close to 0 and every behavioral distribution $\sigma \in \Sigma^*(\hat{\phi},\lambda)$ that society potentially reaches at this $\hat{\phi}$, norm dynamics satisfy: $\dot{\hat{\phi}} < 0$. Thus, $\phi = 0$ is asymptotically stable if $\exists \epsilon > 0$ s.t. $\hat{\phi}(1-\hat{\phi})[(\sigma_1 - \sigma_0)(\gamma v(\hat{\phi}) - \Delta m(\psi^*(\hat{\phi},\lambda))) - \gamma(1-\sigma_1)h + \gamma \Delta k(\hat{\phi})] < 0$ for all $\hat{\phi} \in (0,\epsilon)$ and $\sigma \in \Sigma^*(\hat{\phi},\lambda)$. Since $\hat{\phi}(1-\hat{\phi}) > 0\ \forall \hat{\phi} \notin \{0,1\}$ and $\text{argmin}_a\, \Delta m(a) = 0$, this condition is satisfied if $(\sigma_1 - \sigma_0)(\gamma v(\hat{\phi}) - \Delta m(0)) - \gamma(1-\sigma_1)h + \gamma \Delta k(\hat{\phi}) < 0\ \forall \hat{\phi} \in (0,\epsilon), \sigma \in \Sigma^*(\hat{\phi},\lambda)$. Let ϵ be sufficiently close to 0 s.t. $\gamma \Delta k(\hat{\phi}) < 0$ and $\gamma v(\hat{\phi}) - \Delta m(0) < 0\ \forall \hat{\phi} \in (0,\epsilon)$. Such an ϵ exists due to continuity of Δk and v. Moreover, we know that $(\sigma_1 - \sigma_0) \in [0,1]$ (see Lemma 9) and $\gamma(1-\sigma_1)h \geq 0$. Thus, $(\sigma_1 - \sigma_0)(\gamma v(\hat{\phi}) - \Delta m(0)) - \gamma(1-\sigma_1)h + \gamma \Delta k(0) < 0\ \forall \hat{\phi} \in (0,\epsilon)$ and $\sigma \in \Sigma^*(\hat{\phi},\lambda)$. $\phi = 0$ is an asymptotically stable CE. □

Proof of Proposition 4. Consider any $\lambda \in [0,1]^{|\Theta|}$. $\phi = 1$ is asymptotically stable if for all $\hat{\phi}$ close to 1 and every behavioral distribution $\sigma \in \Sigma^*(\hat{\phi},\lambda)$ that society potentially reaches at this $\hat{\phi}$, norm dynamics satisfy: $\dot{\hat{\phi}} > 0$. Thus, $\phi = 1$ is asymptotically stable if $\exists \epsilon > 0$ s.t. $\hat{\phi}(1-\hat{\phi})[(\sigma_1 - \sigma_0)(\gamma v(\hat{\phi}) - \Delta m(\psi^*(\hat{\phi},\lambda))) - \gamma(1-\sigma_1)h + \gamma \Delta k(\hat{\phi})] > 0$ for all $\hat{\phi} \in (1-\epsilon,1)$ and $\sigma \in \Sigma^*(\hat{\phi},\lambda)$. Since $\hat{\phi}(1-\hat{\phi}) > 0\ \forall \hat{\phi} \in (0,1)$, this condition is satisfied if $(\sigma_1 - \sigma_0)(\gamma v(\hat{\phi}) - \Delta m(\psi^*(1,\lambda))) - \gamma(1-\sigma_1)h + \gamma \Delta k(\hat{\phi}) > 0$ for all $\hat{\phi} \in (1-\epsilon,1)$ and $\sigma \in \Sigma^*(\hat{\phi},\lambda)$.

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

Next, suppose $\Delta k(1) > (1-\psi^*(1,\lambda))h$. ($\Delta k(1) > (1-\psi^*(1,\lambda))h$ and $\lim_{x\to 1}(\sigma_1^*) = \psi^*(1,\lambda)) \Rightarrow \lim_{x\to 1}(\min_{\sigma \in \Sigma^*(x,\lambda)}(\Delta k(x) - (1-\sigma_1)h)) = \Delta k(1) - (1-\psi^*(1,\lambda))h > 0$. Moreover, $\lim_{x\to 1}(\sigma_1^*) = \psi^*(1,\lambda) \Rightarrow \lim_{x\to 1}(\min_{\sigma \in \Sigma^*(x,\lambda)}(\sigma_1(\gamma v(x) - \Delta m(\psi^*(x,\lambda))) - \gamma(1-\sigma_1)h + \gamma \Delta k(x))) = \psi^*(1,\lambda)(\gamma v(1) - \Delta m(\psi^*(1,\lambda))) - \gamma(1-\psi^*(1,\lambda))h + \gamma \Delta k(1) > 0$. Hence, there is some $\epsilon > 0$ s.t. $\forall \hat{\phi} \in (1-\epsilon,1)$ and $\sigma \in \Sigma^*(\hat{\phi},\lambda)$, $-\gamma(1-\sigma_1)h + \gamma \Delta k(\hat{\phi}) > 0$ and $\sigma_1(\gamma v(\hat{\phi}) - \Delta m(\psi^*(\hat{\phi},\lambda))) - \gamma(1-\sigma_1)h + \gamma \Delta k(\hat{\phi})$. Since $(\sigma_1 - \sigma_0) \in [0,\sigma_1]$ (see Lemma 9), $(\sigma_1 - \sigma_0)(\gamma v(\hat{\phi}) - \Delta m(\psi^*(\hat{\phi},\lambda))) - \gamma(1-\sigma_1)h + \gamma \Delta k(\hat{\phi}) > 0$.

Thus, the stated conditions imply that $\phi = 1$ is an asymptotically stable CE at λ. □

Proof of Proposition 5. Consider any $\lambda \in [0,1]^{|\Theta|}$ s.t. $\phi^* = 1$ is a CE of Proposition 4. First, consider the case of $\psi^*(1,\lambda)(\gamma v(1) - \Delta m(\psi^*(1,\lambda))) - \gamma(1-\psi^*(1,\lambda))h + \gamma \Delta k(1) > 0$ and $\Delta k(1) > (1-\psi^*(1,\lambda))h$. Since these inequalities are strict, there is some $\epsilon > 0$ s.t. for all $x \in (\psi^*(1,\lambda) - \epsilon, \psi^*(1,\lambda) + \epsilon)$, $x(\gamma v(1) - \Delta m(x)) - \gamma(1-x)h + \gamma \Delta k(1) > 0$ and $\Delta k(1) > (1-x)h$. Consider any such ϵ. Lemma 16 implies that there is a neighborhood U of λ s.t. $\hat{\lambda} \in U \Rightarrow \psi^*(1,\hat{\lambda}) \in (\psi^*(1,\lambda) - \epsilon, \psi^*(1,\lambda) + \epsilon)$. Hence, there is a neighborhood U of λ s.t. $\hat{\lambda} \in U$ implies that the sufficient conditions for a perfect-social-norm CE are satisfied at $\hat{\lambda}$. Thus, $\phi^* = 1$ is a CE for all $\hat{\lambda} \in U$ for some neighborhood U of λ.

Next, consider the case of $\psi^*(1,\lambda)(\gamma v(1) - \Delta m(\psi^*(1,\lambda))) - \gamma(1-\psi^*(1,\lambda))h + \gamma \Delta k(1) > 0$ and $\theta_s \tilde{v}(1) < \Delta m(\psi^*(1,\lambda))$ for all $\theta \in \text{supp}(\lambda)$. Recall from the proof of Proposition 4 that $\lim_{x\to 1}(\sigma_0^*) = 0$ and $\lim_{x\to 1}(\sigma_1^*) = \psi^*(1,\lambda)$ at λ. For any $\hat{\lambda} \in [0,1]^{|\Theta|}$, $\phi = 1$ is a CE if for all x in some neighborhood of 1 society reaches a behavioral distribution $\sigma \in \Sigma^*(x,\hat{\lambda})$ s.t. $(\sigma_1 - \sigma_0)(\gamma v(x) - \Delta m(\psi^*(x,\hat{\lambda}))) - \gamma(1-\sigma_1)h + \gamma \Delta k(x) > 0$. This holds if $\lim_{x\to 1}(\min_{\sigma \in \Sigma^*(x,\hat{\lambda})}((\sigma_1 - \sigma_0)(\gamma v(x) - \Delta m(\psi^*(x,\hat{\lambda}))) - \gamma(1-\sigma_1)h + \gamma \Delta k(x))) = \lim_{x\to 1}(\min_{\sigma \in \Sigma^*(x,\hat{\lambda})}((\psi^*(1,\hat{\lambda}) - \sigma_0)(\gamma v(1) - \Delta m(\psi^*(1,\hat{\lambda}))) - \gamma(1-\psi^*(1,\hat{\lambda}))h + \gamma \Delta k(1))) > 0$. There is $\alpha > 0$ s.t. for all $y_1 \in (\psi^*(1,\lambda) - \alpha, \psi^*(1,\lambda) + \alpha)$ and $y_0 \in [0,\alpha)$, $(y_1 - y_0)(\gamma v(1) - \Delta m(y_1)) - \gamma(1-y_1)h + \gamma \Delta k(1) > 0$. Consider any such α. Let $\xi \in (0,\alpha)$ be s.t. $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta \in \Theta} |\lambda_\theta - x_\theta| < \xi\} \Rightarrow (\psi^*(1,\hat{\lambda}) \in (\psi^*(1,\lambda) - \alpha, \psi^*(1,\lambda) + \alpha)$ and $\Delta m(\psi^*(1,\hat{\lambda})) > \tilde{v}(1)\forall \theta \in \text{supp}(\lambda))$. Such ξ exists by Lemma 16 and continuity of all involved functions. $\Delta m(\psi^*(1,\hat{\lambda})) > \tilde{v}(1)\forall \theta \in \text{supp}(\lambda) \Rightarrow \exists \epsilon > 0$ s.t. $\Delta m(\psi^*(x,\hat{\lambda})) > \tilde{v}(x)\forall x \in (1-\epsilon,1), \theta \in \text{supp}(\lambda) \Rightarrow \sigma_{0,\theta} = 0\,\forall \theta \in \text{supp}(\lambda), x \in (1-\epsilon,1), \sigma \in \Sigma^*(x,\hat{\lambda})$. Consider any such ϵ. It follows that for all $\theta \in \text{supp}(\lambda), x \in (1-\epsilon,1)$, and $\sigma \in \Sigma^*(x,\hat{\lambda})$, $\sigma_0 = \sum_{\theta \in \text{supp}(\hat{\lambda})} \hat{\lambda}_\theta \sigma_{0,\theta} = \sum_{\theta \in \Theta}(\hat{\lambda}_\theta - \lambda_\theta)\sigma_{0,\theta} + \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \sigma_{0,\theta} = \sum_{\theta \in \Theta}(\hat{\lambda}_\theta - \lambda_\theta)\sigma_{0,\theta} \leq \sum_{\theta \in \Theta}(\hat{\lambda}_\theta - \lambda_\theta) \leq \xi$. Thus, $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta \in \Theta} |\lambda_\theta - x_\theta| < \xi\} \Rightarrow \sigma_0 \leq \alpha\ \forall x \in (1-\epsilon,1), \sigma \in \Sigma^*(x,\hat{\lambda}) \Rightarrow \lim_{x\to 1}(\max_{\sigma \in \Sigma^*(x,\hat{\lambda})}(\sigma_0) \leq \alpha$.

$(\lim_{x\to 1}\max_{\sigma\in\Sigma^*(x,\hat{\lambda})}(\sigma_0)<\alpha$ and $\psi^*(1,\hat{\lambda})\in(\psi^*(1,\lambda)-\alpha,\psi^*(1,\lambda)+\alpha))\Rightarrow\lim_{x\to 1}(\min_{\sigma\in\Sigma^*(x,\hat{\lambda})}((\psi^*(1,\hat{\lambda})-\sigma_0)(\gamma v(1)-\Delta m(\psi^*(1,\hat{\lambda})))-\gamma(1-\psi^*(1,\hat{\lambda}))h+\gamma\Delta k(1)))>0\Rightarrow\phi=1$ is an asymptotically stable CE at $\hat{\lambda}$. □

Proposition 17. *Consider any specification of the model with $\Delta k(\cdot)$, $v(\cdot)$, $\Delta m(\cdot)$, h, γ, and λ s.t. $\text{supp}(\lambda)=\{\theta\}\wedge(\theta_s,\theta_p)<(\gamma(1+\delta),\gamma(h-(1+\delta)\tilde{h}))$.*

1. • *For all $x\in\mathbb{R}_{\geq 0}$, ($\gamma=x\Rightarrow\phi^*=1$ is a CE of Proposition 4) $\Rightarrow$ ($\gamma>x\Rightarrow\phi^*=1$ is a CE of Proposition 4).*
   • *For all $x_1(z)$, ($\Delta k(z)=x(z)\ \forall z\in[0,1]\Rightarrow\phi^*=1$ is a CE of Proposition 4) $\Rightarrow$ ($\Delta k(z)>x(z)\ \forall z\in[0,1]\Rightarrow\phi^*=1$ is a CE of Proposition 4).*
   • *For all $x(z)$, ($v(z)=x(z)\ \forall z\in[0,1]\Rightarrow\phi^*=1$ is a CE of Proposition 4) $\Rightarrow$ ($v(z)>x(z)\ \forall z\in[0,1]\Rightarrow\phi^*=1$ is a CE of Proposition 4).*
   • *For all $x(z)$, ($\Delta m(z)=x(z)\ \forall z\in[0,1]\Rightarrow\phi^*=1$ is a CE of Proposition 4) $\Rightarrow$ ($\Delta m(z)<x(z)\ \forall z\in[0,1]\Rightarrow\phi^*=1$ is a CE of Proposition 4).*
2. (a) • *$\exists x$ s.t. $h<x\Rightarrow\phi^*=1$ is a CE of Proposition 4.*
   • *$\exists x(z)$ s.t. $\Delta m(z)<x(z)\ \forall z\in[0,1]\Rightarrow\phi^*=1$ is a CE of Proposition 4.*
   (b) • *$\exists x(z)$ s.t. $\Delta k(z)>x(z)\ \forall z\in[0,1]\Rightarrow\phi^*=1$ is a CE of Proposition 4.*
   • *$\exists x(z)$ s.t. $v(z)>x(z)\ \forall z\in[0,1]\Rightarrow\phi^*=1$ is a CE of Proposition 4.*
3. $\frac{\partial(z(\gamma v(1)-\Delta m(z))-\gamma(1-z)h+\gamma\Delta k(1))}{\partial z}>0\Leftrightarrow\frac{h+\gamma v(1)-\Delta m(z)}{z}>\Delta m'(z)$.

Proof. First, note that if $\text{supp}(\lambda)$ is a singleton, then $\theta_s\tilde{v}(1)\geq\Delta m(\psi^*(1,\lambda))\Rightarrow\theta_s\tilde{v}(1)+\theta_s\tilde{h}+\theta_p>\Delta m(\psi^*(1,\lambda))\Rightarrow\psi^*(1,\lambda)=1\Rightarrow\Delta k(1)>(1-\psi^*(1,\lambda))h$. Hence, 2a does not hold implies 2b holds. To investigate a perfect-social-norm CE at homogeneous preference distribution λ it, thus, suffices to investigate when Condition 1 of Proposition 4 holds.

Consider any specification of the model. We start with the first statement. Consider any x s.t. $\gamma=x\Rightarrow$ Condition 1 of Proposition 4 holds. Hence, $\psi^*(1,\lambda)(xv(1)-\Delta m(\psi^*(1,\lambda)))-x(1-\psi^*(1,\lambda))h+x\Delta k(1)>0$. $-\Delta m(\psi^*(1,\lambda))<0\Rightarrow\psi^*v(1)-(1-\psi^*(1,\lambda))h+\Delta k(1)>0\Rightarrow(\gamma>x$ and $\Delta m(\cdot),\psi^*(1,\lambda),\Delta k(\cdot),v(\cdot)$ independent of $\gamma\Rightarrow\psi^*(1,\lambda)(\gamma v(1)-\Delta m(\psi^*(1,\lambda)))-\gamma(1-\psi^*(1,\lambda))h+\gamma\Delta k(1)>0\Rightarrow$ Condition 1 of Proposition 4 holds). The proof of the second bullet point of the first statement (concerning $\Delta k(\cdot)$) closely follows the above given the observation that $\Delta k(z)>x(z)\ \forall z\in[0,1]\Rightarrow\Delta k(1)>x(1)$.

Next, consider any $x(z)$ s.t. $v(z)=x(z)\ \forall z\in[0,1]\Rightarrow$ Condition 1 of Proposition 4 holds. Let $\bar{\psi}$ be s.t. $v(z)=x(z)(\Leftrightarrow\tilde{v}(z)=\frac{x(z)}{1+\delta})\ \forall z\in[0,1]\Rightarrow\psi^*(1,\lambda)=\bar{\psi}$. Hence, $\bar{\psi}(\gamma x(1)-\Delta m(\bar{\psi}))-\gamma(1-\bar{\psi})h+\gamma\Delta k(1)>0$. First, suppose $\bar{\psi}=1$. Hence, $\gamma x(1)-\Delta m(1)+\gamma\Delta k(1)>0$. ($v(z)>x(z)\ \forall z\in[0,1]\Rightarrow\tilde{v}(1)>\frac{x(1)}{1+\delta}\Rightarrow\psi^*(1,\lambda)\geq\bar{\psi}\Rightarrow\psi^*(1,\lambda)=1$) and $\gamma v(1)-\Delta m(1)+\gamma\Delta k(1)>0\Rightarrow$ Condition 1 of Proposition 4 holds. Next, suppose $\bar{\psi}\in(0,1)$. Hence, $\bar{\psi}(\gamma x(1)-\theta_s\frac{x(1)}{1+\delta}-\theta_s\tilde{h}-\theta_p)-\gamma(1-\bar{\psi})h+\gamma\Delta k(1)>0\Leftrightarrow\bar{\psi}(\gamma-\frac{\theta_s}{1+\delta})x(1)+\bar{\psi}(\gamma h-\theta_s\tilde{h}-\theta_p)-\gamma h+\gamma\Delta k(1)>0$. $v(z)>x(z)\ \forall z\in[0,1]\Rightarrow\tilde{v}(1)>\frac{x(1)}{1+\delta}\Rightarrow\psi^*(1,\lambda)\geq\bar{\psi}$. $v(1)>x(1)$, $\psi^*(1,\lambda)\geq\bar{\psi}$, $\gamma(1+\delta)>\theta_s$, and $\theta_p+\theta_s\tilde{h}<\theta_p+\gamma(1+\delta)\tilde{h}<\gamma h\Rightarrow\psi^*(1,\lambda)(\gamma-\frac{\theta_s}{1+\delta})v(1)+\psi^*(1,\lambda)(\gamma h-\theta_s\tilde{h}-\theta_p)-\gamma h+\gamma\Delta k(1)>0$ imply Condition 1 of Proposition 4 holds. The proof of the last bullet point of the first statement (concerning $\Delta m(\cdot)$) closely resembles the previous one.

Next, we turn to the second statement. First, consider $x<\Delta k(1)$. $h<x\Rightarrow$ (a) $\psi^*(1,\lambda)=0\wedge\Delta k(1)>h\Rightarrow$ Condition 1 of Proposition 4 holds, (b) $\psi^*(1,\lambda)\in(0,1)\Rightarrow\Delta m(\psi^*(1,\lambda))=\theta_s\tilde{v}(1)+\theta_s\tilde{h}+\theta_p\Rightarrow(\psi^*(1,\lambda)(\gamma v(1)-\theta_s\tilde{v}(1)-\theta_s\tilde{h}-\theta_p)-\gamma(1-\psi^*(1,\lambda))h+\gamma\Delta k(1)>0\Leftrightarrow\psi^*(1,\lambda)(\gamma-\frac{\theta_s}{1+\delta})v(1)+\psi^*(1,\lambda)(\gamma h-\theta_s\tilde{h}-\theta_p)-\gamma h+\gamma\Delta k(1)>0$, $\Delta k(1)>(1-\psi^*(1,\lambda))h>h$, $\theta_s<(1+\delta)\gamma$, $\theta_s\tilde{h}+\theta_p<\gamma h)\Rightarrow$ Condition 1 of Proposition 4 holds, (c) $\psi^*(1,\lambda)=1\Rightarrow\theta_s\tilde{v}(1)+\theta_s\tilde{h}+\theta_p\geq\Delta m(1)\Rightarrow\gamma v(1)+\gamma h>\Delta m(1)\Rightarrow\gamma v(1)+\gamma\Delta k(1)>\Delta m(1)\Rightarrow$ Condition 1 of Proposition 4 holds.

Next, consider any $x(z)$ s.t. $\gamma x(1)>\Delta m(1)+\gamma h$. $\Delta k(z)>x(z)\ \forall z\in[0,1]\Rightarrow\Delta k(1)>x(1)$. $(\psi^*(1,\lambda)(\gamma v(1)-\Delta m(\psi^*(1,\lambda)))-\gamma(1-\psi^*(1,\lambda))h+x(1)>-\Delta m(1)-\gamma h+x(1)>0)\Rightarrow(\Delta k(1)>x(1)\Rightarrow$ Condition 1 of Proposition 4 holds).

Next, consider any $x(z)$ s.t. $\Delta m(1)<\min\{\gamma x(1)+\gamma\Delta k(1),\theta_s\frac{x(1)}{1+\delta}+\theta_s\tilde{h}+\theta_p\}$. $v(1)>x(1)\Rightarrow\psi^*(1,\lambda^d)=1\wedge\Delta m(1)<\gamma v(1)+\gamma\Delta k(1)\Rightarrow$ Condition 1 of Proposition 4 holds.

Next, consider any $x(z)$ s.t. $x(1)<\min\{\gamma v(1)+\gamma\Delta k(1),\theta_s\tilde{v}(1)+\theta_s\tilde{h}+\theta_p\}$. By the same reasoning as above, Conditions 1 and 2 of Proposition 4 hold.

Lastly, statement 3 follows from taking the partial derivative of Condition 1 of Proposition 4. □

Proof of Proposition 7. Consider any $\lambda\in[0,1]^{|\Theta|}$ and $\phi\in(0,1)$ s.t. the stated conditions hold. $\theta_s\tilde{v}(\phi)<\Delta m(\phi)<\theta_s\tilde{v}(\phi)+\theta_s\tilde{h}+\theta_p\ \forall\theta\in\text{supp}(\lambda)\Rightarrow(\sigma_1,\sigma_0)=(1,0)\ \forall\sigma\in\Sigma^*(\phi,\lambda)$. Moreover, $\exists\epsilon>0$ s.t. $\theta_s\tilde{v}(\hat{\phi})<\Delta m(\hat{\phi})<\theta_s\tilde{v}(\hat{\phi})+\theta_s\tilde{h}+\theta_p\ \forall\theta\in\text{supp}(\lambda)$ and $\hat{\phi}\in(\phi-\epsilon,\phi+\epsilon)\Rightarrow(\sigma_1,\sigma_0)=(1,0)$ for all $\hat{\phi}\in(\phi-\epsilon,\phi+\epsilon)$ and $\sigma\in\Sigma^*(\hat{\phi},\lambda)$. Thus, the equilibrium values σ_0^*, σ_1^*, and $\psi^*(\phi,\lambda)$ are continuous and differentiable at ϕ; $\frac{d\sigma_0^*}{dx}|_{x=\phi}=0$, $\frac{d\sigma_1^*}{dx}|_{x=\phi}=0$, and $\frac{\partial\psi^*(x,\lambda)}{\partial x}|_{x=\phi}=1$.

Consider norm dynamics (Definition 7). $((\sigma_1,\sigma_0)=(1,0)$ and $\gamma(v(\phi^*)+\Delta k(\phi^*))=\Delta m(\phi^*))\Rightarrow\dot{\phi}=0$. Thus, ϕ is a rest point. $\gamma(\frac{dv(x)}{dx}|_{x=\phi^*}+\frac{d\Delta k(x)}{dx}|_{x=\phi})<\frac{d\Delta m(\phi)}{dx}|_{x=\phi}$ ensures asymptotic stability. To see this, note that since equilibrium behavior satisfies $(\sigma_1^*,\sigma_0^*)=(1,0)$ in some interval around ϕ, we can write norm dynamics at ϕ as a function of the social norm only. Moreover, $\dot{\hat{\phi}}(1-\hat{\phi})>0\ \forall\hat{\phi}\in(\phi-\epsilon,\phi+\epsilon)$. Thus, a rest point is asymptotically stable if

$$\frac{d[C_1(\sigma^*,x)-C_0(\sigma^*,x)]}{dx}|_{x=\phi}=\gamma(\frac{dv(x)}{dx}|_{x=\phi}+\frac{d\Delta k(x)}{dx}|_{x=\phi})-\frac{d\Delta m(x)}{dx}|_{x=\phi}<0.$$

Consequently, ϕ is asymptotically stable under the stated conditions and, thus, an asymptotically stable CE. □

Proposition 18. *Consider any $\lambda \in [0,1]^{|\Theta|}$ for which a CE $\phi^* \in (0,1)$ exists. For all $\epsilon > 0$, there is $\xi > 0$ s.t. $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta} |\lambda_\theta - x_\theta| < \xi\}$ implies that there is some $\hat{\Phi} \subset (\phi^* - \epsilon, \phi^* + \epsilon)$ s.t.*

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

Proof. Consider any $\lambda \in [0,1]^{|\Theta|}$ for which some CE $\phi^* \in (0,1)$ exists. Moreover, consider some $\epsilon > 0$. Let $\eta \in (0,\epsilon)$ be s.t. (1) $\dot{x} > 0$ ($\Leftrightarrow (\sigma_1 - \sigma_0)(\gamma v(x) - \Delta m(x\sigma_1 + (1-x)\sigma_0) - \gamma(1-\sigma_1)h + \gamma\Delta k(x) > 0$) for all $x \in [\phi - \eta, \phi)$ and (2) $\dot{x} < 0$ ($\Leftrightarrow (\sigma_1 - \sigma_0)(\gamma v(x) - \Delta m(x\sigma_1 + (1-x)\sigma_0) - \gamma(1-\sigma_1)h + \gamma\Delta k(x) < 0$) for all $x \in (\phi, \phi + \eta]$ and for all $\sigma \in \Sigma^*(x,\lambda)$. Such an η exists since ϕ is a CE at λ. Let $\overline{x} = \phi + \eta$, and $\underline{x} = \phi - \eta$.

Consider σ_0 and σ_1 for all $\sigma \in \Sigma^*(\underline{x},\lambda)$. Since Δm, $\tilde{v}$, and $\tilde{h}$ are continuous, $\exists \alpha > 0$ s.t. $|\sigma_n - \hat{\sigma}_n| < \alpha\, \forall n \in \{0,1\} \Rightarrow (\hat{\sigma}_1 - \hat{\sigma}_0)(\gamma v(\underline{x}) - \Delta m(\underline{x}\hat{\sigma}_1 + (1-\underline{x})\hat{\sigma}_0)) - \gamma(1-\hat{\sigma}_1)\tilde{h} + \gamma\Delta k(\underline{x}) > 0$. Consider any such $\alpha > 0$. Proposition 17 implies that $\exists \underline{\xi}$ s.t. $\forall \hat{\lambda} \in [0,1]^{|\Theta|}$, $\sum_{\theta\in\Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \underline{\xi}$ implies that for all $\hat{\sigma} \in \Sigma^*(\underline{x},\hat{\lambda})$ and $\sigma \in \Sigma^*(\underline{x},\lambda)$, $|\sigma_n - \hat{\sigma}_n| < \alpha\, \forall n \in \{0,1\}$. Consequently, $\sum_{\theta\in\Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \underline{\xi} \Rightarrow \dot{\underline{x}} > 0$ at any $\hat{\sigma} \in \Sigma^*(\underline{x},\hat{\lambda})$.

Analogously, we can show that there is some $\overline{\xi} > 0$ s.t. for all $\hat{\lambda} \in [0,1]^{|\Theta|}$, $\sum_{\theta\in\Theta} |\lambda_\theta - \hat{\lambda}_\theta| < \overline{\xi}$ implies that $\dot{\overline{x}} < 0$ at any $\hat{\sigma} \in \Sigma^*(\overline{x},\hat{\lambda})$.

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

Next, we investigate norm evolution at any other $\check{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta} |\lambda_\theta - x_\theta| < \xi\}$, when starting at some element $\hat{\phi} \in \hat{\Phi}$. $\hat{\phi} \in \hat{\Phi} \subset (\underline{x}, \overline{x})$ and $\dot{\underline{x}} > 0 \wedge \dot{\overline{x}} < 0$ at $\check{\lambda}$ imply that there is some minimal, asymptotically stable set $\check{\Phi} \subset (\underline{x}, \overline{x})$ that norms evolve to when starting at $\hat{\phi} \in \hat{\Phi}$. □

Lemma 18. *Consider any $\lambda \in [0,1]^{|\Theta|}$ for which a CE $\phi^* \in (0,1)$ with equilibrium share $\psi^*(\phi^*,\lambda) = \phi^*$ exists. Moreover, let $\tau \in \mathbb{R}_{\geq 0}$ be s.t. $\tau v(\phi^*) < \Delta m(\phi^*) < \tau h + \tau v(\phi^*)$. There is $\xi > 0$ s.t. $\hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta} |\lambda_\theta - x_\theta| < \xi\}$ implies that $\tau v(\hat{\phi}) < \Delta m(\psi^*(\hat{\phi},\hat{\lambda})) < \tau h + \tau v(\hat{\phi})$ for all $\hat{\phi} \in \hat{\Phi}^*$, where $\hat{\Phi}^*$ is the minimal, asymptotically stable set at $\hat{\lambda}$ with ϕ^* in it's basin of attraction.*

Proof. Consider any $\lambda \in [0,1]^{|\Theta|}$ for which a CE $\phi^* \in (0,1)$ with equilibrium share $\psi^*(\phi^*,\lambda) = \phi^*$ exists and $\tau \in \mathbb{R}_{\geq 0}$ s.t. $\tau v(\phi^*) < \Delta m(\phi^*) < \tau h + \tau v(\phi^*)$. Consider any $\alpha, \beta > 0$ s.t. $\forall x \in (\phi^* - \alpha, \phi^* + \alpha), y \in (\phi^* - \beta, \phi^* + \beta)$, $\tau v(x) < \Delta m(y) < \tau h + \tau v(x)$. Such α and β exists due to continuity of all involved functions. Consider any $\bar{\beta} < \beta$ so small that for all $\phi \in (\phi^* - \bar{\beta}, \phi^* + \bar{\beta})$, $\psi^*(\phi,\lambda) \in (\phi^* - \beta, \phi^* + \beta)$. Lemma 15 implies that such $\bar{\beta}$ exists. Let $\xi > 0$ be s.t. $\forall \hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta} |\lambda_\theta - x_\theta| < \xi\}$, $\hat{\phi} \in (\phi^* - \min\{\alpha, \check{\beta}\}, \phi^* + \min\{\alpha, \check{\beta}\})$ for all $\hat{\phi} \in \hat{\Phi}^*$, where $\hat{\Phi}^*$ is the minimal, asymptotically stable set at $\hat{\lambda}$ with ϕ^* in it's basin of attraction. Such ξ exists due to Proposition 18. Consequently, for all $\forall \hat{\lambda} \in \{x \in [0,1]^{|\Theta|} : \sum_{\theta\in\Theta} |\lambda_\theta - x_\theta| < \xi\}$ and $\hat{\phi} \in \hat{\Phi}^*$, $\tau v(\hat{\phi}) < \Delta m(\psi^*(\hat{\phi},\hat{\lambda})) < \tau h + \tau v(\hat{\phi})$. □

Proof of Lemma 1. Consider any λ s.t. $\text{supp}(\lambda) = \{\theta\} \wedge \theta < (\gamma(1+\delta), \gamma(h - (1+\delta)\tilde{h}))$. We start with the first statement. Consider any $z_1 < z_2$ and $I_y = \{\phi \in [0,1] : \theta_s \tilde{v}(\phi) < \gamma v(\phi) + \gamma\Delta k(\phi) < \theta_s \tilde{v}(\phi) + z_y\}\ \forall y \in \{1,2\}$. To proof the statement, it suffices to show that $I_1 \subseteq I_2$, which is clearly true.

Next, consider the second statement of the lemma. Suppose $\theta_s \tilde{h} + \theta_p = 0$. Hence, $I_p(\lambda) = \{\phi \in [0,1] : \theta_s \tilde{v}(\phi) < \gamma v(\phi) + \gamma\Delta k(\phi) < \theta_s \tilde{v}(\phi)\} = \emptyset$. Lastly, it remains for us to show that $\theta_s \tilde{h} + \theta_p > 0 \Rightarrow I_p(\lambda) \neq \emptyset$. Note that $\gamma v(0) + \gamma\Delta k(0) = \gamma(1+\delta)\tilde{v}(0) + \gamma\Delta k(0) < \theta_s \tilde{v}(0)$. $\theta_s > (1+\delta)\gamma \wedge \Delta k(\frac{1}{2}) = 0 \Rightarrow \gamma v(\frac{1}{2}) + \gamma\Delta k(\frac{1}{2}) = \gamma(1+\delta)\tilde{v}(\frac{1}{2}) + \gamma\Delta k(\frac{1}{2}) > \theta_s \tilde{v}(\frac{1}{2})$. Hence, $\exists x \in (0, \frac{1}{2})$ s.t. (i) $\theta_s \tilde{v}(x) = \gamma v(x) + \gamma\Delta k(x) < \theta_s \tilde{v}(x) + \theta_s \tilde{h} + \theta_p$ and $\theta_s \tilde{v}'(x) < \gamma v'(x) + \gamma\Delta k'(x)$. Thus, $\exists \epsilon \in (0, \frac{1}{2} - x)$ s.t. $\forall y \in (x, x+\epsilon)$, $\theta_s \tilde{v}(y) = \gamma v(y) + \gamma\Delta k(y) < \theta_s \tilde{v}(y) + \theta_s \tilde{h} + \theta_p$. Hence, $I_p(\lambda) \cap (0, \frac{1}{2}) \neq \emptyset \Rightarrow I_p(\lambda) \neq \emptyset$. □

Proposition 19. *Consider any specification with $\Delta k(\cdot)$, $v(\cdot)$, $\Delta m(\cdot)$, γ and λ s.t. $\text{supp}(\lambda) = \{\theta\}$. Consider any $\phi \in (0,1)$.*

1. • *For all $x(z)$, ($v(z) = x(z)\ \forall z \in [0,1] \Rightarrow \phi \in (0,1)$ is a CE of Proposition 7 at λ) $\Rightarrow$ ($\exists \epsilon > 0$ s.t. $v(z) \in (x(z), x(z) + \epsilon)\ \forall z \in [0,1] \Rightarrow \exists \Phi$ s.t. $\phi < \min(\Phi)$ and Φ is a minimal, asymptotically stable set at λ).*
   • *For all $x(z)$, ($\Delta m(z) = x(z)\ \forall z \in [0,1] \Rightarrow \phi \in (0,1)$ is a CE of Proposition 7 at λ) $\Rightarrow$ ($\exists \epsilon > 0$ s.t. $\Delta m(z) \in (x(z) - \epsilon, x(z))\ \forall z \in [0,1] \Rightarrow \exists \Phi$ s.t. $\phi < \min(\Phi)$ and Φ is a minimal, asymptotically stable set at λ).*
   • *For all $x \in \mathbb{R}_{\geq 0}$, ($\gamma = x \Rightarrow \phi \in (0,1)$ is a CE of Proposition 7 at λ) $\Rightarrow$ ($\exists \epsilon > 0$ s.t. $\gamma \in (x, x+\epsilon) \Rightarrow \exists \Phi$ s.t. $\phi < \min(\Phi)$ and Φ is a minimal, asymptotically stable set at λ).*
2. *For all $x(z)$, ($\Delta k(z) = x(z)\ \forall z \in [0,1] \Rightarrow \phi \in (0,1)$ is a CE of Proposition 7 at λ) $\Rightarrow$ ($|\Delta k(z)| > |x(z)|\ \forall z \in [0,1] \Rightarrow \exists \Phi$ s.t. Φ is a CE at λ and (a) $\phi < \min(\Phi)$ if $\frac{1}{2} < \phi$ and (b) $\phi > \max(\Phi)$ if $\frac{1}{2} > \phi$).*

Proof. Consider any specification of the model with $\Delta k(\cdot)$, $v(\cdot)$, $\Delta m(\cdot)$, γ, δ, λ s.t. $\text{supp}(\lambda) = \{\theta\}$, and any $\phi \in (0,1)$.

For statement 1 of the proposition, suppose some $x(z)$ s.t. $v(z) = x(z)\ \forall z \in [0,1] \Rightarrow \phi \in (0,1)$ is a CE of Proposition 7 at λ, which implies $\gamma x(\phi) + \gamma\Delta k(\phi) = \Delta m(\phi)$ and $\theta_s \frac{x(\phi)}{1+\delta} < \Delta m(\phi) < \theta_s \frac{x(\phi)}{1+\delta} + \theta_s \tilde{h} + \theta_p$. Let $\epsilon < \frac{\Delta m(\phi)}{\theta_s} - \frac{x(z)}{1+\delta}$. $v(z) \in (x(z), x(z) + \epsilon)\ \forall z \in [0,1] \Rightarrow$

$v(\phi) \in (x(\phi), x(\phi)+\epsilon) \Rightarrow (\gamma v(\phi) + \gamma \Delta k(\phi) > \Delta m(\phi) \wedge \theta_s \tilde{v}(\phi) < \Delta m(\phi) < \theta_s \tilde{v}(\phi) + \theta_s \tilde{h} + \theta_p) \Rightarrow \dot{\phi} > 0$. Hence, $v(z) \in (x(z), x(z)+\epsilon)\ \forall z \in [0,1]$ implies that $\dot{\phi} > 0$, implying that the norm dynamics move to some minimal, asymptotically stable set Φ s.t. $\phi < \min(\Phi)$. The proof for the remaining bullet points of statement 1 work analogously.

The proof of statement 2 works analogously to the above with the preceding observation that $|\Delta k(z)| > |x(z)|\ \forall z \in [0,1] \Rightarrow |\Delta k(\phi)| > |x(\phi)| \Rightarrow$ (a) $\Delta k(\phi) > x(\phi)$ if $\phi > \frac{1}{2}$ and (b) $\Delta k(\phi) < x(\phi)$ if $\phi < \frac{1}{2}$. □

B.3. Approval preferences

Lemma 19. *For all* $\lambda \in \{x \in [0,1]^{|\Theta|} : \theta^d \in \text{supp}(x)\}, \theta \in \text{supp}(\lambda), \phi \in [0,1], n \in \{0,1\}$, *and* $\sigma \in \Sigma^*(\phi, \lambda)$, $B_{n,\theta^d}(\sigma, \phi) \geq B_{n,\theta}(\sigma, \phi)$ *and* $B_{\theta^d}(\sigma, \phi) \geq B_\theta(\sigma, \phi)$.

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

Proof of Lemma 2. Consider any $\lambda \in [0,1]^{|\Theta|}$ s.t. $\phi^* \in [0,1]$ is a CE. Hence, norm dynamics (Definition 7) are at rest, $\dot{\phi}^* = 0$. The two conditions in Lemma 2 can be jointly expressed as $\sigma_n = \bar{\sigma}_n\ \forall n \in \{0,1\} \setminus \{1-\phi^*\}, \sigma \in \Sigma^*(\phi^*, \lambda^d)$, and $\bar{\sigma} \in \Sigma^*(\phi^*, \lambda)$. Moreover, $(\sigma_1, \sigma_0) = (\bar{\sigma}_1, \bar{\sigma}_0) \Rightarrow \psi^*(\phi^*, \lambda^d) = \psi^*(\phi^*, \lambda)$. Throughout, we write $\psi^* := \psi^*(\phi^*, \lambda) = \psi^*(\phi^*, \lambda^d)$.

Consider any $n \in \{0,1\} \setminus \{1-\phi^*\}$. First, we investigate the case s.t. $\exists a \in \{0,1\}$ s.t. $b(a, n, \psi^*, \phi^*) > b(1-a, n, \psi^*, \phi)$. For all $\theta \in \text{supp}(\lambda)$, $\sigma \in \Sigma^*(\phi^*, \lambda^d)$, and $\bar{\sigma} \in \Sigma^*(\phi^*, \lambda)$, $b(a, n, \psi^*, \phi^*) > b(1-a, n, \psi^*, \phi^*) \Rightarrow \sigma_n = a = \bar{\sigma}_n$. $\bar{\sigma}_n = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \bar{\sigma}_{n,\theta} = a \in \{0,1\} \Rightarrow \bar{\sigma}_{n,\theta} = a\ \forall \theta \in \text{supp}(\lambda) \Rightarrow B_{n,\hat{\theta}} = b(a, n, \psi^*, \phi^*) = B_{n,\tilde{\theta}}\ \forall \tilde{\theta}, \hat{\theta} \in \text{supp}(\lambda)$. Next, we look at the case of $b(0, n, \psi^*, \phi^*) = b(1, n, \psi^*, \phi^*)$. $b(0, n, \psi^*, \phi^*) = b(1, n, \psi^*, \phi^*) \Rightarrow (1-y) b(0, n, \psi^*, \phi^*) + y b(1, n, \psi^*, \phi^*) = (1-x) b(0, n, \psi^*, \phi^*) + x b(1, n, \psi^*, \phi^*) \forall x, y \in [0,1] \Rightarrow B_{n,\hat{\theta}} = B_{n,\tilde{\theta}} \forall \tilde{\theta}, \hat{\theta} \in \text{supp}(\lambda)$. Thus, for all $\lambda \in [0,1]^{|\Theta|}, \phi^* \in [0,1], n \in \{0,1\} \setminus \{1-\phi^*\}, \sigma \in \Sigma^*(\phi^*, \lambda^d)$, and $\bar{\sigma} \in \Sigma^*(\phi^*, \lambda)$, $\sigma_n = \bar{\sigma}_n \Rightarrow B_{n,\hat{\theta}}(\bar{\sigma}, \phi^*) = B_{n,\tilde{\theta}}(\bar{\sigma}, \phi^*)$. Hence, $\dot{\lambda}_\theta = 0\ \forall \theta \in \text{supp}(\lambda)$. For all $\theta \notin \text{supp}(\lambda)$, $\dot{\lambda}_\theta = 0$. Hence, (λ, ϕ^*) is a rest point of the dynamic system. □

Proof of Proposition 10. Proposition 3 implies that $\phi^* = 0$ is an asymptotically stable CE at any $\lambda \in [0,1]^{|\Theta|}$. Hence, in some neighborhood of $(\lambda, 0)$, society always coordinates to $\phi^* = 0$ before changes in preferences occur. Moreover, $\tilde{v}(0) = 0 < \Delta m(0)\ \forall \theta \in \Theta \Rightarrow \psi^*(0, \lambda) = 0\ \forall \lambda \in [0,1]^{|\Theta|}$. $\forall \lambda \in [0,1]^{|\Theta|}$, $(\lambda, 0)$ is a rest point. Consequently, $(\lambda, 0)$ is stable $\forall \lambda \in [0,1]^{|\Theta|}$. □

Proof of Proposition 11. Consider any $\lambda \in [0,1]^{|\Theta|}$ s.t. $\psi^*(1, \lambda) = \psi^*(1, \lambda^d)$ and suppose $\phi^* = 1$ is a CE of Proposition 4 at λ. Lemma 2 implies that $(\lambda, 1)$ is a rest point.

Consider any U of λ s.t. $\phi^* = 1$ is a CE for all $\hat{\lambda} \in U$. Such a U exists by Proposition 5. In some neighborhood of $(\lambda, 1)$, society always coordinates to $\phi^* = 1$ before changes in preferences occur. Therefore, we can reduce our attention to preference dynamics.

First, suppose $\psi^*(1, \lambda) = \psi^*(1, \hat{\lambda}) = \psi^*(1, \lambda^d)$. $\psi^*(1, \lambda^d) = x \in \{0,1\} \Rightarrow \hat{\sigma}_{1,\theta} = x\ \forall \theta \in \text{supp}(\hat{\lambda}) \Rightarrow B_{\hat{\lambda}}(\hat{\sigma}, 1) = \sum_{\theta \in \text{supp}(\hat{\lambda})} \hat{\lambda}_\theta (\hat{\sigma}_{1,\theta} b(1, 1, x, 1) + (1 - \hat{\sigma}_{1,\theta}) b(0, 1, x, 1)) = b(x, 1, x, 1) = B_\lambda(\hat{\sigma}, 1) = \sum_{\theta \in \text{supp}(\hat{\lambda})} \lambda_\theta (\hat{\sigma}_{1,\theta} b(1, 1, x, 1) + (1 - \hat{\sigma}_{1,\theta}) b(0, 1, x, 1))\ \forall \hat{\sigma} \in \Sigma^*(1, \hat{\lambda}) \Rightarrow B_{\hat{\lambda}}(\hat{\sigma}, 1) = B_\lambda(\hat{\sigma}, 1)\ \forall \hat{\sigma} \in \Sigma^*(1, \hat{\lambda})$. $\psi^*(1, \lambda^d) \in (0,1) \Rightarrow b(1, 1, \psi^*(1, \lambda^d), 1) = b(0, 1, \psi^*(1, \lambda^d), 1) \Rightarrow B_{\hat{\lambda}}(\hat{\sigma}, 1) = B_\lambda(\hat{\sigma}, 1)\ \forall \hat{\sigma} \in \Sigma^*(1, \hat{\lambda})$.

Next, suppose $\psi^*(1, \hat{\lambda}) < \psi^*(1, \lambda) = 1$. $\psi^*(1, \lambda) = 1 \Rightarrow \psi^*(1, \lambda^d) = 1 \Rightarrow \theta^d_s \tilde{v}(1) + \theta^d_s \tilde{h} + \theta^d_p \geq \Delta m(1) \Rightarrow b(1, 1, 1, 1) \geq b(0, 1, 1, 1)$. $\psi^*(1, \hat{\lambda}) < 1 \Rightarrow \Delta m(\psi^*(1, \hat{\lambda})) < \Delta m(1) \Rightarrow b(1, 1, \psi^*(1, \hat{\lambda}), 1) - b(0, 1, \psi^*(1, \hat{\lambda}), 1) > b(1, 1, 1, 1) - b(0, 1, 1, 1) \geq 0$. Moreover, $\psi^*(1, \lambda) = 1 \Rightarrow \theta_p + \theta_s \tilde{h} + \theta_s \tilde{v}(1) \geq \Delta m(1) \forall \theta \in \text{supp}(\lambda) \Rightarrow \theta_p + \theta_s \tilde{h} + \theta_s \tilde{v}(1) > \Delta m(\psi^*(1, \hat{\lambda})) \forall \theta \in \text{supp}(\lambda) \Rightarrow \hat{\sigma}_{1,\theta} = 1 \forall \theta \in \text{supp}(\lambda), \hat{\sigma} \in \Sigma^*(1, \hat{\lambda}) \Rightarrow B_\lambda(\hat{\sigma}, 1) = b(1, 1, \psi^*(1, \hat{\lambda}), 1)\ \forall \hat{\sigma} \in \Sigma^*(1, \hat{\lambda})$. Therefore, $B_\lambda(\hat{\sigma}, 1) = b(1, 1, \psi^*(1, \hat{\lambda}), 1) > b(1, 1, \psi^*(1, \hat{\lambda}), 1) \psi^*(1, \hat{\lambda}) + (1 - \psi^*(1, \hat{\lambda})) b(0, 1, \psi^*(1, \hat{\lambda}), 1) = B_{\hat{\lambda}}(\hat{\sigma}, 1)\ \forall \hat{\sigma} \in \Sigma^*(1, \hat{\lambda})$.

Analog to the previous case, we can show that $\psi^*(1, \hat{\lambda}) > \psi^*(1, \lambda) = 0$ implies $B_\lambda(\hat{\sigma}, 1) > B_{\hat{\lambda}}(\hat{\sigma}, 1) \forall \hat{\sigma} \in \Sigma^*(1, \hat{\lambda})$. We refrain form writing it out.

Next, suppose $\phi^* = 1 > \psi^*(1, \lambda) > \psi^*(1, \hat{\lambda}) > 0$. $\psi^*(1, \lambda) \in (0,1) \Rightarrow \psi^*(1, \lambda^d) \in (0,1) \Rightarrow \theta^d_s \tilde{v}(1) + \theta^d_s \tilde{h} + \theta^d_p = \Delta m(\psi^*(1, \lambda^d)) \Rightarrow b(1, 1, \psi^*(1, \lambda), 1) = b(0, 1, \psi^*(1, \lambda), 1)$. $\psi^*(1, \hat{\lambda}) < \psi^*(1, \lambda) \Rightarrow \Delta m(\psi^*(1, \hat{\lambda})) < \Delta m(\psi^*(1, \lambda)) \Rightarrow b(1, 1, \psi^*(1, \hat{\lambda}), 1) - b(0, 1, \psi^*(1, \hat{\lambda}), 1) > b(1, 1, \psi^*(1, \lambda), 1) - b(0, 1, \psi^*(1, \lambda), 1) = 0$. Consider any $\sigma \in \Sigma^*(1, \lambda), \hat{\sigma} \in \Sigma^*(1, \hat{\lambda})$ and let $\hat{\sigma}_{1,\lambda} = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \hat{\sigma}_{1,\theta}$. $\Delta m(\psi^*(1, \hat{\lambda})) < \Delta m(\psi^*(1, \lambda)) \Rightarrow (\forall \theta \in \Theta, \theta_p + \theta_s \tilde{h} + \theta_s \tilde{v}(1) \geq \Delta m(\psi^*(1, \lambda)) \Rightarrow \theta_p + \theta_s \tilde{h} + \theta_s \tilde{v}(1) > \Delta m(\psi^*(1, \hat{\lambda}))) \Rightarrow \hat{\sigma}_{1,\theta} \geq \sigma_{1,\theta}\ \forall \theta \in \text{supp}(\lambda) \Rightarrow \hat{\sigma}_{1,\lambda} = \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \hat{\sigma}_{1,\theta} \geq \sum_{\theta \in \text{supp}(\lambda)} \lambda_\theta \sigma_{1,\theta} = \psi^*(1, \lambda) > \psi^*(1, \hat{\lambda})$. ($b(1, 1, \psi^*(1, \hat{\lambda}), 1) > b(0, 1, \psi^*(1, \hat{\lambda}), 1)$ and $\hat{\sigma}_{1,\lambda} > \psi^*(1, \hat{\lambda})$) $\Rightarrow B_\lambda(\hat{\sigma}, 1) = \hat{\sigma}_{1,\lambda} b(1, 1, \psi^*(1, \hat{\lambda}), 1) + (1 - \hat{\sigma}_{1,\lambda}) b(0, 1, \psi^*(1, \hat{\lambda}), 1) > \psi^*(1, \hat{\lambda}) b(1, 1, \psi^*(1, \hat{\lambda}), 1) + (1 - \psi^*(1, \hat{\lambda})) b(0, 1, \psi^*(1, \hat{\lambda}), 1) = B_{\hat{\lambda}}(\hat{\sigma}, 1)$. Hence, $B_\lambda(\hat{\sigma}, 1) > B_{\hat{\lambda}}(\hat{\sigma}, 1)$.

Lastly, the case of $\phi^* = 1 > \psi^*(1, \hat{\lambda}) > \psi^*(1, \lambda) > 0$ works analogously to the previous one. Therefore, we refrain from writing it out.

Hence, for all different cases of $\psi^*(1,\lambda)$ and $\psi^*(1,\hat{\lambda})$, $B_\lambda(\hat{\sigma},1) \geq B_{\hat{\lambda}}(\hat{\sigma},1)\ \forall \hat{\sigma} \in \Sigma^*(1,\hat{\lambda})$. Any mutation from λ to $\hat{\lambda} \in U$ does not alter $\phi^* = 1$. For any post-mutation NE $\hat{\sigma} \in \Sigma^*(1,\hat{\lambda})$, $B_\lambda(\hat{\sigma},1) \geq B_{\hat{\lambda}}(\hat{\sigma},1)$. Following Weibull (1997), this condition ensures that λ is stable in preference dynamics (Definition 9). Hence, $(\lambda,1)$ is stable. □

Proposition 20. *Consider any specification of the model s.t.* $\psi^*(1,\lambda^d) \in (0,1)$.

1. $\frac{\mathrm{d}\Delta m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)} < \frac{(v(1)+h)(\gamma-\rho)}{\psi^*(1,\lambda^d)} \Rightarrow \frac{\mathrm{d}\psi^*(1,\lambda^d)(\gamma v(1)-\Delta m(\psi^*(1,\lambda^d)))}{\mathrm{d}\rho} - \frac{\mathrm{d}\gamma(1-\psi^*(1,\lambda^d))h}{\mathrm{d}\rho} + \frac{\mathrm{d}\gamma\Delta k(1)}{\mathrm{d}\rho} > 0.$
2. $\frac{\mathrm{d}\Delta m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)} < \frac{(v(1)+h)(\gamma-\rho)\rho}{\gamma-\psi^*(1,\lambda^d)(\gamma-\rho)} \Rightarrow \frac{\mathrm{d}\psi^*(1,\lambda^d)(\gamma v(1)-\Delta m(\psi^*(1,\lambda^d)))}{\mathrm{d}h} - \frac{\mathrm{d}\gamma(1-\psi^*(1,\lambda^d))h}{\mathrm{d}h} + \frac{\mathrm{d}\gamma\Delta k(1)}{\mathrm{d}h} > 0.$

Proof. We start with the first statement. $\psi^*(1,\lambda) \in (0,1) \Rightarrow \Delta m(0) < \Delta m(\psi^*(1,\lambda^d)) = \theta_s^d \tilde{v}(1) + \theta_s^d \tilde{h} + \theta_p^d < \Delta m(1)$. For small changes in ρ, it remains that $\Delta m(0) < \Delta m(\psi^*(1,\lambda)) = \theta_s^d \tilde{v}(1) + \theta_s^d \tilde{h} + \theta_p^d < \Delta m(1)$. Hence, $\frac{\mathrm{d}\Delta m(\psi^*(1,\lambda))}{\mathrm{d}\rho} = \frac{\mathrm{d}(\theta_s^d \tilde{v}(1)+\theta_s^d \tilde{h}+\theta_p^d)}{\mathrm{d}\rho}$, from which we can derive $\frac{\mathrm{d}\Delta m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)} * \frac{\mathrm{d}\psi^*(1,\lambda^d)}{\mathrm{d}\rho} = h + v(1)$. $\frac{\mathrm{d}\psi^*(1,\lambda^d)(\gamma v(1)-\Delta m(\psi^*(1,\lambda^d)))}{\mathrm{d}\rho} - \frac{\mathrm{d}\gamma(1-\psi^*(1,\lambda^d))h}{\mathrm{d}\rho} + \frac{\mathrm{d}\gamma\Delta k(1)}{\mathrm{d}\rho} > 0 \Leftrightarrow \frac{\mathrm{d}\psi^*(1,\lambda^d)(\gamma-\rho)(v(1)+h)}{\mathrm{d}\rho} > 0 \Leftrightarrow \frac{\mathrm{d}\psi^*(1,\lambda^d)}{\mathrm{d}\rho} > \frac{\psi^*(1,\lambda^d)}{\gamma-\rho} \Leftrightarrow \frac{(\gamma-\rho)(v(1)+h)}{\psi^*(1,\lambda^d)} > \frac{\mathrm{d}\Delta m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)}$. Hence, statement 1 of the proposition is true.

Next, consider the second statement of the proposition. By similar reasoning as above, we can show that $\frac{\mathrm{d}\Delta m(\psi^*(1,\lambda))}{\mathrm{d}h} = \frac{\mathrm{d}(\theta_s^d \tilde{v}(1)+\theta_s^d \tilde{h}+\theta_p^d)}{\mathrm{d}h}$, which implies $\frac{\mathrm{d}\Delta m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)} * \frac{\mathrm{d}\psi^*(1,\lambda^d)}{\mathrm{d}h} = \rho$. $\frac{\mathrm{d}\psi^*(1,\lambda^d)(\gamma v(1)-\Delta m(\psi^*(1,\lambda^d)))}{\mathrm{d}h} - \frac{\mathrm{d}\gamma(1-\psi^*(1,\lambda^d))h}{\mathrm{d}h} + \frac{\mathrm{d}\gamma\Delta k(1)}{\mathrm{d}h} > 0 \Leftrightarrow \frac{\mathrm{d}\psi^*(1,\lambda^d)(\gamma-\rho)(v(1)+h)}{\mathrm{d}h} - \frac{\mathrm{d}\gamma h}{\mathrm{d}h} > 0 \Leftrightarrow (\gamma-\rho)(v(1)+h)\frac{\mathrm{d}\psi^*(1,\lambda^d)}{\mathrm{d}h} + \psi^*(1,\lambda^d)(\gamma-\rho) - \gamma > 0 \Leftrightarrow \frac{(\gamma-\rho)(v(1)+h)\rho}{\gamma-\psi^*(1,\lambda^d)(\gamma-\rho)} > \frac{\mathrm{d}\Delta m(x)}{\mathrm{d}x}|_{x=\psi^*(1,\lambda^d)}$. Hence, the second statement of the proposition is also true. □

Proof of Proposition 13. Consider any CE $\phi^* = 1$ at λ^d s.t. Conditions 1 and 2b of Proposition 4 hold. Hence,

1. $\psi^*(1,\lambda^d)(\gamma v(1) - \Delta m(\psi^*(1,\lambda^d))) - \gamma(1-\psi^*(1,\lambda^d))h + \gamma\Delta k(1) > 0$ and
2. $\Delta k(1) > (1-\psi^*(1,\lambda^d))h$.

Throughout, let $\Lambda := \{(1,\lambda) : \psi^*(1,\lambda) = \psi^*(1,\lambda^d)\}$. Consider any $\lambda \in \Lambda$. Λ is non-empty since λ^d is always in it. $\lambda \in \Lambda \Rightarrow \psi^*(1,\lambda^d) = \psi^*(1,\lambda)$. Hence, the two conditions above also hold for λ and the perfect social norm is a CE of Proposition 4 at λ.

Consider any U of λ s.t. $\phi^* = 1$ is a CE for all $\hat{\lambda} \in U$. Such a U exists by Proposition 5. $\hat{\lambda} \in U \setminus \Lambda \Rightarrow \phi^* = 1$ is a CE at $\hat{\lambda}$ and $\psi^*(1,\hat{\lambda}) \neq \psi^*(1,\lambda)$. By the same reasoning as in the proof of Proposition 11, we can show that $\psi^*(1,\hat{\lambda}) \neq \psi^*(1,\lambda) \Rightarrow B_\lambda(\hat{\sigma},1) > B_{\hat{\lambda}}(\hat{\sigma},1)\ \forall \hat{\sigma} \in \Sigma^*(1,\hat{\lambda})$. The mutation from preference distribution $\lambda \in \Lambda$ to $\hat{\lambda} \in U \setminus \Lambda$ does not alter the perfect social norm $\phi^* = 1$. For any post-mutation NE $\hat{\sigma} \in \Sigma^*(1,\hat{\lambda})$, $B_\lambda(\hat{\sigma},1) > B_{\hat{\lambda}}(\hat{\sigma},1)$. Following Weibull (1997), this condition ensures that approval preferences evolve towards some $\check{\lambda} \in \Lambda$ and, thus, return to the set Λ. Throughout the course of preference evolution, the perfect social norm remains a CE. The proposition is true. □

Proof of Proposition 14. Consider any CE $\phi^* \in (0,1)$ of Proposition 7 at λ^d. $(\sigma_1,\sigma_0) = (1,0)\ \forall \sigma \in \Sigma^*(\phi^*,\lambda^d) \Rightarrow b(n,n,\phi^*,\phi^*) > b(1-n,n,\phi^*,\phi^*)\ \forall n \in \{0,1\}$. Moreover, consider any λ s.t. $\phi^* \in (0,1)$ is a CE of Proposition 7 at λ. Hence $\theta_s\tilde{v}(\phi^*) < \Delta m(\phi^*) < \theta_s\tilde{v}(\phi^*) + \theta_s\tilde{h} + \theta_p\ \forall \theta \in \mathrm{supp}(\lambda)$ and $(\sigma_1,\sigma_0) = (1,0)\ \forall \sigma \in \Sigma^*(\phi^*,\lambda)$. Lemma 2 implies (λ,ϕ^*) is a rest point.

For any $(\hat{\lambda},\hat{\phi})$ close to (λ,ϕ^*), we write the CE that society reaches at $(\hat{\lambda},\hat{\phi})$ as $\hat{\phi}^*$. We continue to show that there is some U of (λ,ϕ^*) s.t. for all $(\hat{\lambda},\hat{\phi}) \in U$, $B_\lambda(\hat{\sigma},\hat{\phi}^*) > B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi}^*)\ \forall \hat{\sigma} \in \Sigma^*(\hat{\phi}^*,\hat{\lambda})$. Let the neighborhood U of (λ,ϕ^*) be s.t. for all $(\hat{\lambda},\hat{\phi}) \in U$:

1. $\Sigma^*(\hat{\phi}^*,\hat{\lambda})$ is a singleton for all $\hat{\phi}^* \neq \phi^*$,
2. $\theta_s\tilde{v}(\phi^*) < \Delta m(\phi^*) < \theta_s\tilde{v}(\phi^*) + \theta_s\tilde{h} + \theta_p \Rightarrow \theta_s\tilde{v}(\hat{\phi}^*) < \Delta m(\psi^*(\hat{\phi}^*,\hat{\lambda})) < \theta_s\tilde{v}(\hat{\phi}^*) + \theta_s\tilde{h} + \theta_p$ for all $\theta \in \Theta$,
3. $b(n,n,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*) > b(1-n,n,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*)\ \forall n \in \{0,1\}$, and
4. $\hat{\phi}^*$ is in the basin of attraction of ϕ^* at preference distribution λ and λ^d.

Such U exists due to Lemma 12, Proposition 18, and Lemma 18.

Note, $\hat{\phi}^* \neq \phi^* \Rightarrow (\hat{\sigma}_1,\hat{\sigma}_0) \neq (1,0)\ \forall \hat{\sigma} \in \Sigma^*(\hat{\phi}^*,\hat{\lambda})$. Assume by contradiction that $\hat{\phi}^* \neq \phi^*$ and $(\hat{\sigma},\hat{\sigma}_0) = (1,0)$ for $\hat{\sigma} \in \Sigma^*(\hat{\phi}^*,\hat{\lambda})$. Recall from the above that $\Sigma^*(\hat{\phi}^*,\hat{\lambda})$ is a singleton. At λ^d, (a) $\hat{\phi}^*$ is in the basin of attraction of ϕ^* and (b) $(\sigma_1,\sigma_0) = (1,0)$ for all $\sigma \in \Sigma^*(\hat{\phi}^*,\lambda^d)$. Hence, $(\sigma_1,\sigma_0) = (1,0)$ and $\hat{\phi}^* \neq 0$ for all $\sigma \in \Sigma^*(\hat{\phi}^*,\lambda^d)$. If $(\hat{\sigma}_1,\hat{\sigma}_0) = (1,0)$ for all $\hat{\sigma} \in \Sigma^*(\hat{\phi}^*,\hat{\lambda})$, then $\hat{\phi}^* \neq 0$ must also hold at preference distribution $\hat{\lambda}$. Consequently, $\hat{\phi}^*$ is not a rest point, which is a contradiction.

Consider any $\hat{\sigma} \in \Sigma^*(\hat{\phi}^*,\hat{\lambda})$. First, suppose $(\hat{\sigma}_1,\hat{\sigma}_0) \neq (1,0)$. $\theta_s\tilde{v}(\phi^*) < \Delta m(\phi^*) < \theta_s\tilde{v}(\phi^*) + \theta_s\tilde{h} + \theta_p\ \forall \theta \in \mathrm{supp}(\lambda) \Rightarrow \theta_s\tilde{v}(\hat{\phi}^*) < \Delta m(\psi^*(\hat{\phi}^*,\hat{\lambda})) < \theta_s\tilde{v}(\hat{\phi}^*) + \theta_s\tilde{h} + \theta_p\ \forall \theta \in \mathrm{supp}(\lambda) \Rightarrow \hat{\sigma}_{n,\theta} = n\ \forall \theta \in \mathrm{supp}(\lambda)$ and $n \in \{0,1\}$. Hence, $B_\lambda(\hat{\sigma},\hat{\phi}^*) = \hat{\phi}^* b(1,1,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*) + (1-\hat{\phi}^*)b(0,0,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*) \wedge B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi}^*) = \hat{\phi}^*[\hat{\sigma}_1 b(1,1,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*) + (1-\hat{\sigma}_1)b(0,1,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*)] + (1-\hat{\phi}^*)[\hat{\sigma}_0 b(1,0,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*) + (1-\hat{\sigma}_0)b(0,0,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*)]] \Rightarrow B_\lambda(\hat{\sigma},\hat{\phi}^*) - B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi}^*) = \hat{\phi}^*(1-\hat{\sigma}_1)(b(1,1,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*) - b(0,1,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*)) + (1-\hat{\phi}^*)\hat{\sigma}_0(b(0,0,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*) - b(1,0,\psi^*(\hat{\phi}^*,\hat{\lambda}),\hat{\phi}^*)) > 0 \Rightarrow B_\lambda(\hat{\sigma},\hat{\phi}^*) > B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi}^*)$. Second, suppose $(\hat{\sigma}_1,\hat{\sigma}_0) = (1,0)$, which implies $\hat{\phi}^* = \phi^*$. $\hat{\sigma}_n = n\ \forall n \in \{0,1\} \Rightarrow \hat{\sigma}_{n,\tilde{\theta}} = \hat{\sigma}_{n,\check{\theta}} = n\ \forall n \in \{0,1\}, \tilde{\theta}, \check{\theta} \in \mathrm{supp}(\hat{\lambda}) \Rightarrow B_\lambda(\hat{\sigma},\hat{\phi}^*) = B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi}^*)$.

At $(\hat{\lambda},\hat{\phi}) \in U$, society first coordinates into a CE $\hat{\phi}^*$. At $(\hat{\lambda},\hat{\phi}^*)$ and any $\hat{\sigma} \in \Sigma^*(\hat{\phi}^*,\hat{\lambda})$, $B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi}^*) \geq B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi}^*)$. Following Weibull (1997), λ is stable on preference dynamics. Throughout the course of preference evolution, the CE that society reach remain very close to ϕ^* (see Proposition 18). Hence, Proposition 14 is true. □

Proof of Proposition 15. If $I_p(\lambda^d)$ is non-empty, then any $m(\cdot)$ that renders conditions 2 and 3 of Proposition 7 true yields the Proposition 15 to be true. If $I_p(\lambda^d)$ is empty, then there is no ϕ s.t. ϕ is a CE of Proposition 7 at λ^d. From Lemma 1, we know that $I_p(\lambda^d)$ is non empty if and only if $\theta_s^d \tilde{h} + \theta_p^d > 0$. Substituting for θ^d yields $I_p(\lambda^d)$ is non-empty if and only if $\rho h > 0$. Since $\rho > 0$, $I_p(\lambda^d)$ is non empty if and only if $h > 0$. □

Proof of Lemma 3. Consider any CE ϕ^* of Proposition 7 at λ^d. Suppose that for all $\theta \in \Theta, n \in \{0,1\}$, $u(1,n,\phi^*,\phi^*) \neq u(1,n,\phi^*,\phi^*)$. Let $\Lambda = \{x \in [0,1]^{|\Theta|} : (\sigma_1,\sigma_0) = (1,0)\ \forall \sigma \in \Sigma^*(\phi^*,x)\}$. Consider any $\lambda \in \Lambda$. $(\sigma_1,\sigma_0) = (1,0)\ \forall \sigma \in \Sigma^*(\phi^*,\lambda)$ and $u(1,n,\phi^*,\phi^*) \neq u(1,n,\phi^*,\phi^*)\ \forall \theta \in \Theta, n \in \{0,1\}$ implies ϕ^* is a CE of Proposition 7 at λ. For any $(\hat{\lambda},\hat{\phi})$ close to (λ,ϕ^*), we write the CE that society reaches at $(\hat{\lambda},\hat{\phi})$ as $\hat{\phi}^*$. Let U of λ be as in the proof of Proposition 14.

Consider any $\hat{\lambda} \in U \setminus \Lambda$. As in the proof of Proposition 14, we can show that $\hat{\phi}^* \neq \phi^* \Rightarrow (\hat{\sigma}_1,\hat{\sigma}_0) \neq (1,0)\ \forall \hat{\sigma} \in \Sigma^*(\hat{\phi}^*,\hat{\lambda})$. In conjunction with $\hat{\lambda} \notin \Lambda \Rightarrow (\hat{\phi}^* = \phi^* \Rightarrow (\hat{\sigma}_1,\hat{\sigma}_0) \neq (1,0))$, it follows that $(\hat{\sigma}_1,\hat{\sigma}_0) \neq (1,0)\ \forall \hat{\sigma} \in \Sigma^*(\hat{\phi}^*,\hat{\lambda})$. By similar reasoning as in the proof of Proposition 14, we can show that $B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi}^*) > B_{\hat{\lambda}}(\hat{\sigma},\hat{\phi}^*)$.

Following Weibull (1997), this condition ensures that preferences evolve towards some $\check{\lambda} \in \Lambda$. Throughout the course of preference evolution, the CE society reaches remains close to ϕ^* (see Proposition 18). Once preferences reach an element in Λ, the social norm returns to ϕ^* implying that the proposition is true. □

References

Abeler, J., Nosenzo, D., Raymond, C., 2019. Preferences for truth-telling. Econometrica 87, 1115–1153. https://doi.org/10.3982/ECTA14673.

Akerlof, G.A., Kranton, R.E., 2005. Identity and the economics of organizations. J. Econ. Perspect. 19, 9–32. https://doi.org/10.1257/0895330053147930.

Alger, I., Weibull, J.W., 2013. Homo moralis—preference evolution under incomplete information and assortative matching. Econometrica 81, 2269–2302. https://doi.org/10.3982/ECTA10637.

Alger, I., Weibull, J.W., 2016. Evolution and kantian morality. Games Econ. Behav. 98, 56–67. https://doi.org/10.1016/j.geb.2016.05.006.

Andreoni, J., Bernheim, B.D., 2009. Social image and the 50–50 norm: a theoretical and experimental analysis of audience effects. Econometrica 77, 1607–1636. https://doi.org/10.3982/ECTA7384.

Azar, O.H., 2004. What sustains social norms and how they evolve?: the case of tipping. J. Econ. Behav. Organ. 54, 49–64. https://doi.org/10.1016/j.jebo.2003.06.001.

Bašić, Z., Quercia, S., 2022. The influence of self and social image concerns on lying. Games Econ. Behav. 133, 162–169. https://doi.org/10.1016/j.geb.2022.02.006.

Bénabou, R., Tirole, J., 2006. Incentives and prosocial behavior. Am. Econ. Rev. 96, 1652–1678. https://doi.org/10.1257/aer.96.5.1652.

Bénabou, R., Tirole, J., 2011. Identity, morals, and taboos: beliefs as assets. Q. J. Econ. 126, 805–855. https://doi.org/10.1093/qje/qjr002.

Bereczkei, T., Csanaky, A., 1996. Mate choice, marital success, and reproduction in a modern society. Ethol. Sociobiol. 17, 17–35. https://doi.org/10.1016/0162-3095(95)00104-2.

Bernheim, B.D., 1994. A theory of conformity. J. Polit. Econ. 102, 841–877. https://doi.org/10.1086/261957.

Bester, H., Güth, W., 1998. Is altruism evolutionarily stable? J. Econ. Behav. Organ. 34, 193–209. https://doi.org/10.1016/S0167-2681(97)00060-7.

Bezin, E., 2019. The economics of green consumption, cultural transmission and sustainable technological change. J. Econ. Theory 181, 497–546. https://doi.org/10.1016/j.jet.2019.03.005.

Binmore, K., Samuelson, L., 1994. An economist's perspective on the evolution of norms. J. Inst. Theor. Econ. 150, 45–63. https://www.jstor.org/stable/40753015.

Bisin, A., Topa, G., Verdier, T., 2004. Cooperation as a transmitted cultural trait. Ration. Soc. 16, 477–507. https://doi.org/10.1177/1043463104046695.

Bisin, A., Verdier, T., 1998. On the cultural transmission of preferences for social status. J. Public Econ. 70, 75–97. https://doi.org/10.1016/S0047-2727(98)00061-9.

Bisin, A., Verdier, T., 2001. The economics of cultural transmission and the dynamics of preferences. J. Econ. Theory 97, 298–319. https://doi.org/10.1006/jeth.2000.2678.

Blau, P.M., 1964. Exchange and Power in Social Life. Transaction Publishers, NJ.

Bowles, S., Gintis, H., 1998. The moral economy of communities: structured populations and the evolution of pro-social norms. Evol. Hum. Behav. 19, 3–25. https://doi.org/10.1016/S1090-5138(98)00015-4.

Boyd, R., Richerson, P.J., 1990. Group selection among alternative evolutionarily stable strategies. J. Theor. Biol. 145, 331–342. https://doi.org/10.1016/S0022-5193(05)80113-4.

Boyd, R., Richerson, P.J., 2005. The Origin and Evolution of Cultures. Oxford University Press.

Brekke, K.A., Kverndokk, S., Nyborg, K., 2003. An economic model of moral motivation. J. Public Econ. 87, 1967–1983. https://doi.org/10.1016/S0047-2727(01)00222-5.

Buss, D.M., Schmitt, D.P., 1993. Sexual strategies theory: an evolutionary perspective on human mating. Psychol. Rev. 100, 204–232. https://doi.org/10.1037/0033-295x.100.2.204.

Carbonara, E., Parisi, F., Von Wangenheim, G., 2008. Lawmakers as norm entrepreneurs. Rev. Law Econ. 4, 779–799. https://doi.org/10.2202/1555-5879.1320.

Chudek, M., Henrich, J., 2011. Culture–gene coevolution, norm-psychology and the emergence of human prosociality. Trends Cogn. Sci. 15, 218–226. https://doi.org/10.1016/j.tics.2011.03.003.

Cinyabuguma, M., Page, T., Putterman, L., 2005. Cooperation under the threat of expulsion in a public goods experiment. J. Public Econ. 89, 1421–1435. https://doi.org/10.1016/j.jpubeco.2004.05.011.

Cooter, R., 1998. Expressive law and economics. J. Leg. Stud. 27, 585–607. https://doi.org/10.1086/468036.

Crawford, S.E.S., Ostrom, E., 1995. A grammar of institutions. Am. Polit. Sci. Rev. 89, 582–600. https://doi.org/10.2307/2082975.

d'Adda, G., Dufwenberg, M., Passarelli, F., Tabellini, G., 2020. Social norms with private values: theory and experiments. Games Econ. Behav. 124, 288–304. https://doi.org/10.1016/j.geb.2020.08.012.

Elster, J., 1989. Social norms and economic theory. J. Econ. Perspect. 3, 99–117. https://doi.org/10.1257/jep.3.4.99.

Fehr, E., Fischbacher, U., 2004. Social norms and human cooperation. Trends Cogn. Sci. 8, 185–190. https://doi.org/10.1016/j.tics.2004.02.007.

Fershtman, C., Weiss, Y., 1998. Social rewards, externalities and stable preferences. J. Public Econ. 70, 53–73. https://doi.org/10.1016/S0047-2727(98)00060-7.

Figueres, C., Masclet, D., Willinger, M., 2013. Weak moral motivation leads to the decline of voluntary contributions. J. Public Econ. Theory 15, 745–772. https://doi.org/10.1111/jpet.12036.

Fisman, R., Kariv, S., Markovits, D., 2007. Individual preferences for giving. Am. Econ. Rev. 97, 1858–1876. https://doi.org/10.1016/S0047-2727(98)00060-7.

Gächter, S., Fehr, E., 1999. Collective action as a social exchange. J. Econ. Behav. Organ. 39, 341–369. https://doi.org/10.1016/S0167-2681(99)00045-1.
Geary, D.C., Vigil, J., Byrd-Craven, J., 2004. Evolution of human mate choice. J. Sex Res. 41, 27–42. https://doi.org/10.1080/00224490409552211.
Gintis, H., 2003a. The hitchhiker's guide to altruism: gene-culture coevolution, and the internalization of norms. J. Theor. Biol. 220, 407–418. https://doi.org/10.1006/jtbi.2003.3104.
Gintis, H., 2003b. Solving the puzzle of prosociality. Ration. Soc. 15, 155–187. https://doi.org/10.1177/1043463103015002001.
Gintis, H., 2011. Gene–culture coevolution and the nature of human sociality. Philos. Trans. - R. Soc. B, Biol. Sci. 366, 878–888. https://doi.org/10.1098/rstb.2010.0310.
Gintis, H., Smith, E.A., Bowles, S., 2001. Costly signaling and cooperation. J. Theor. Biol. 213, 103–119. https://doi.org/10.1006/jtbi.2001.2406.
Güth, W., Yaari, M.E., 1992. An evolutionary approach to explain reciprocal behavior in a simple strategic game. In: Witt, U. (Ed.), Explaining Process and Change: Approaches to Evolutionary Economics. University of Michigan Press, pp. 23–34.
Guttman, J.M., 2003. Repeated interaction and the evolution of preferences for reciprocity. Econ. J. 113, 631–656. https://doi.org/10.1111/1468-0297.t01-1-00144.
Guttman, J.M., 2013. On the evolution of conditional cooperation. Eur. J. Polit. Econ. 30, 15–34. https://doi.org/10.1016/j.ejpoleco.2012.11.003.
Henrich, J., 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. J. Econ. Behav. Organ. 53, 3–35. https://doi.org/10.1016/S0167-2681(03)00094-5.
Henrich, J., Boyd, R., 1998. The evolution of conformist transmission and the emergence of between-group differences. Evol. Hum. Behav. 19, 215–241. https://doi.org/10.1016/S1090-5138(98)00018-X.
Henrich, J., Boyd, R., 2001. Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. J. Theor. Biol. 208, 79–89. https://doi.org/10.1006/jtbi.2000.2202.
Henrich, J., Gil-White, F.J., 2001. The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. Evol. Hum. Behav. 22, 165–196. https://doi.org/10.1016/S1090-5138(00)00071-4.
Hofbauer, J., Sandholm, W.H., 2009. Stable games and their dynamics. J. Econ. Theory 144, 1665–1693. https://doi.org/10.1016/j.jet.2009.01.007.
Irons, W., 1979. Cultural and biological success. In: Chagnon, N., Irons, W. (Eds.), Evolutionary Biology and Human Social Behavior: An Anthropological Perspective. Duxbury Press, North Scituate, MA, pp. 257–272.
Lindbeck, A., Nyberg, S., Weibull, J.W., 1999. Social norms and economic incentives in the welfare state. Q. J. Econ. 114, 1–35. https://doi.org/10.1162/003355399555936.
Mengel, F., 2008. Matching structure and the cultural transmission of social norms. J. Econ. Behav. Organ. 67, 608–623. https://doi.org/10.1016/j.jebo.2008.01.001.
Michaeli, M., Spiro, D., 2015. Norm conformity across societies. J. Public Econ. 132, 51–65. https://doi.org/10.1016/j.jpubeco.2015.09.003.
Mitteldorf, J., Wilson, D.S., 2000. Population viscosity and the evolution of altruism. J. Theor. Biol. 204, 481–496. https://doi.org/10.1006/jtbi.2000.2007.
Müller, S., von Wangenheim, G., 2019. Coevolution of cooperation, preferences, and cooperative signals in social dilemmas. Discussion Paper 221. Center for European, Governance, and Economic Development Research.
Nordblom, K., Žamac, J., 2012. Endogenous norm formation over the life cycle-the case of tax morale. Econ. Anal. Policy 42. https://doi.org/10.1016/S0313-5926(12)50017-2.
Nyborg, K., 2000. Homo economicus and homo politicus: interpretation and aggregation of environmental values. J. Econ. Behav. Organ. 42, 305–322. https://doi.org/10.1016/S0167-2681(00)00091-3.
Nyborg, K., 2018. Social norms and the environment. Annu. Rev. Resour. Econ. 10, 405–423. https://doi.org/10.1146/annurev-resource-100517-023232.
Nyborg, K., Rege, M., 2003a. Does public policy crowd out private contributions to public goods. Public Choice 115, 397–418. https://doi.org/10.1023/A:1024245522958.
Nyborg, K., Rege, M., 2003b. On social norms: the evolution of considerate smoking behavior. J. Econ. Behav. Organ. 52, 323–340. https://doi.org/10.1016/S0167-2681(03)00031-3.
Ostrom, E., 2000. Collective action and the evolution of social norms. J. Econ. Perspect. 14, 137–158. https://doi.org/10.1257/jep.14.3.137.
Panebianco, F., 2016. The role of persuasion in cultural evolution dynamics. Int. Rev. Econ. 63, 233–258. https://doi.org/10.1007/s12232-016-0253-4.
Poulsen, A., Poulsen, O., 2006. Endogenous preferences and social-dilemma institutions. J. Inst. Theor. Econ. 162, 627–660. https://doi.org/10.1628/093245606779252742.
Rabin, M., 1995. Moral Preferences, Moral Constraints, and Self-Serving Biases. Working Paper 95-241. Berkeley Department of Economics.
Rege, M., 2004. Social norms and private provision of public goods. J. Public Econ. Theory 6, 65–77. https://doi.org/10.1111/j.1467-9779.2004.00157.x.
Richerson, P., Boyd, R., 2010. The Darwinian theory of human cultural evolution and gene-culture coevolution. In: Bell, M.A., Futuyma, D.J., Eanes, W.F., Levinton, J.S. (Eds.), Evolution Since Darwin: The First 150 Years. Sinauer, pp. 266–290.
Richerson, P.J., Boyd, R., Henrich, J., 2010. Gene-culture coevolution in the age of genomics. Proc. Natl. Acad. Sci. 107, 8985–8992. https://doi.org/10.1073/pnas.0914631107.
Sandholm, W.H., 2010. Population Games and Evolutionary Dynamics. MIT Press.
Sethi, R., Somanathan, E., 1996. The evolution of social norms in common property resource use. Am. Econ. Rev. 86, 766–788. https://www.jstor.org/stable/2118304.
Shackelford, T.K., Schmitt, D.P., Buss, D.M., 2005. Universal dimensions of human mate preferences. Pers. Individ. Differ. 39, 447–458. https://doi.org/10.1016/j.paid.2005.01.023.
Tabellini, G., 2008. The scope of cooperation: values and incentives. Q. J. Econ. 123, 905–950. https://doi.org/10.1162/qjec.2008.123.3.905.
Thøgersen, J., 2006. Norms for environmentally responsible behaviour: an extended taxonomy. J. Environ. Psychol. 26, 247–261. https://doi.org/10.1016/j.jenvp.2006.09.004.
Traxler, C., 2010. Social norms and conditional cooperative taxpayers. Eur. J. Polit. Econ. 26, 89–103. https://doi.org/10.1016/j.ejpoleco.2009.11.001.
Traxler, C., Spichtig, M., 2011. Social norms and the indirect evolution of conditional cooperation. J. Econ. 102, 237–262. https://doi.org/10.1007/s00712-010-0173-9.
Turke, P.W., 1989. Evolution and the demand for children. Popul. Dev. Rev. 15, 61–90. https://doi.org/10.2307/1973405.
Voss, T., 2001. Game theoretical perspectives on the emergence of social norms. In: Hechter, M., Opp, K.D. (Eds.), Social Norms. Russell Sage Foundation, pp. 105–138.
Weibull, J.W., 1997. Evolutionary Game Theory. MIT Press.
Wiederman, M.W., 1993. Evolved gender differences in mate preferences: evidence from personal advertisements. Ethol. Sociobiol. 14, 331–351. https://doi.org/10.1016/0162-3095(93)90003-Z.
Young, H.P., 1993. The evolution of conventions. Econometrica 61, 57–84. https://doi.org/10.2307/2951778.
Young, H.P., 1996. The economics of convention. J. Econ. Perspect. 10, 105–122. https://doi.org/10.1257/jep.10.2.105.
Young, H.P., 2015. The evolution of social norms. Annu. Rev. Econ. 7, 359–387. https://doi.org/10.1146/annurev-economics-080614-115322.