WORKING PAPERS

N° 1605

December 2024

“Norms and norm change - driven by social preferences and Kantian morality”

Ingela Alger and Péter Bayer

Toulouse School of Economics

Norms and norm change — driven by social preferences and Kantian morality[*]

Ingela Alger[†] Péter Bayer[‡]

December 31, 2024

Abstract

Norms indicate which behaviors are commonly expected and/or considered to be morally right. We examine how such norms come about and change by modeling a population of individuals with preferences – found elsewhere to be evolutionarily founded – combining material self-interest, Kantian moral concerns, and attitudes towards being materially ahead and behind others. The individuals interact in a public goods game. We identify conditions on preferences and beliefs which promote, respectively hamper, spontaneous norm change. Crucially, an individual's preferences and beliefs about the material benefits uniquely determines her threshold for collective behavior: s/he contributes if and only if sufficiently many others do so. However, those with sufficiently strong Kantian concerns contribute regardless.

Keywords: moral norms, descriptive norms, social norms, social-Kantian preferences

[*] I.A. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789111 - ERC EvolvingEconomics) and both authors acknowledge IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program).

[†] Toulouse School of Economics, CNRS, University of Toulouse Capitole, and Institute for Advanced Study in Toulouse, France, and CEPR. ingela.alger@tse-fr.eu

[‡] Universitàt Autonoma Barcelona and Barcelona School of Economics, Spain. peter.bayer@uab.cat

1 Introduction

Washing one's hands every now and then, stepping outside to smoke at a party, throwing trash in a garbage bin rather than on the ground, picking up one's dog's leavings, respecting the order of arrival in queues; these are but a few examples of behaviors whereby positive externalities are generated at a cost to their authors. In many countries most of these behaviors are nowadays considered natural, in the sense that deviations trigger both surprise and disapproval. However, such natural behaviors, or norms, may evolve over time and vary across space. To wit, a few decades ago smokers could indulge in their habit indoors, and queuing behaviors are not the same everywhere.

We propose a theoretical model which can explain norm variation over time and space. In this it shares the objective of a large number of existing theoretical models on social norms, conventions, and other behavioral regularities; for early treatments and recent surveys see, e.g., Ullman-Margarit (1977), Schelling (1978), Elster (1989), Bicchieri (1990, 2006), Young (1993, 2015), Binmore (1998), Lindbeck et al. (1999), Nyborg & Rege (2003), Nyborg (2018), Bicchieri, Muldoon, & Sontuoso (2023), and Gavrilets et al. (2024).[1] Its contribution consists in showing that a specific preference class, which differs qualitatively from preference classes explored in the literature, can fruitfully be brought to bear on this question. In particular, this preference class is sufficient to generate spontaneous changes in norms, for easily graspable and intuitive reasons, and fully in line with rational choice. It recently emerged from the theoretical analysis on the evolutionary foundations of preferences (Alger et al. (2020)). It has also been found to enhance the explanatory power for behavior in lab experiments (Miettinen et al. (2020), van Leeuwen & Alger (2024)), and we will use the preference parameter estimates of van Leeuwen & Alger (2024) to illustrate our theoretical results.

The said preference class, henceforth social-Kantian preferences, combines three motivations: material self-interest, attitudes towards being ahead and being behind others materially (Fehr & Schmidt (1999)), and a Kantian moral concern (Alger & Weibull (2013)). The Kantian moral concern endogenizes the *personal moral norms* that individuals form – what they deem is the right thing to do. Specifically, an individual with a Kantian moral concern decides on this by applying a

[1] Norms are objects of research in many disciplines, and inter-disciplinary approaches seem appealing; for recent contributions on their potential, see, for example, Bicchieri, Dimant, et al. (2023), Alger et al. (2024), Andrighetto et al. (2024), Gelfand et al. (2024), and Heyes (2024).

universalization argument, evaluating each course of action in the light of the material payoff they would obtain if – hypothetically – the others were also to select this course of action. The social concerns make individuals care about whether they make a greater or a smaller material sacrifice than others, and hence their beliefs about others' behavior – their *empirical expectations* – matter. Our model differs from earlier literature on two key aspects: (1) models which do include personal moral norms take them to be exogenously given; see Gavrilets (2021) as well as Gavrilets et al. (2024) and references therein;[2] (2) the literature has focused on a desire to conform with others' behaviors, without regard to differences in material well-being that a different behavior entails; see, e.g., Bernheim (1994), Glaeser & Scheinkman (2000), Brock & Durlauf (2001), Blume & Durlauf (2003), Bisin et al. (2006), Blume et al. (2015), and Arduini et al. (2022).[3] In the literature there is thus no clear link between the material payoff consequences and norms and norm change. By contrast, with social-Kantian preferences, once an individual's weights on the social and Kantian motivations, her beliefs about the material payoff consequences of her actions, and her empirical expectations are known, one can predict both her personal moral norm and how she evaluates deviations from following the action adopted by others.

We model a population of N individuals, all equipped with social-Kantian preferences, though the weights they attach to the Kantian and the social motivations may be different (they may even be nil). Each individual's preference type is given and fixed through time; we do not examine the evolution of preferences in the population, only the evolution of behaviors. The context is a linear public goods game with two actions in which, from a material perspective, not contributing is the privately rational action: the marginal material benefit to self falls short of the cost. Our main focus is on settings where the collectively rational action is to contribute, i.e., the marginal societal impact exceeds the individual cost of contributing. While individuals with some Kantian concern form their personal norm by considering which action would maximize their material payoff if all individuals were to adopt it, individuals without any Kantian concerns do not form such personal norms: they only care about their actual material payoff and possibly also about how it compares

[2] Models in which personal moral norms are endogenous, such as, for example Brekke et al. (2003) (discussed also by Nyborg (2018)) and López-Pérez (2008), do not analyze norm dynamics.

[3] Our model further rules out reputational or image concerns (see, e.g., Kuran (1989), Akerlof (1997), Gavrilets (2020), and te Velde (2022)), punishment (see, e.g., Hauert et al. (2007), and Richerson et al. (2021)), and identity (see, e.g., Akerlof & Kranton (2000) and Kuran & Sandholm (2008)). We do not claim that these forces are irrelevant; rather, we propose a complementary mechanism behind norms and norm change.

to that of the others. The two main questions we address are: What is the set of possible behavioral norms in this population, where a behavioral norm is the most commonly observed equilibrium behavior? Once a behavioral norm is established, which factors can make it change? And which factors favor an alignment between behavioral norms and personal moral norms?

Prior to describing our results, it seems important to relate our model to norm concepts found in the literature. We follow the literature by referring to what an individual deems is "the right thing to do" as her personal, or private, moral norm (Elster (1989), Cialdini & Goldstein (2004),Bicchieri (2006)), and "the expectation of what others in one's reference network do" as the empirical expectation (Bicchieri, Muldoon, & Sontuoso (2023)). By contrast, our concept of "behavioral norm" differs from conventions, which may arise in pure coordination situations, such as whether left or right is applied when driving or holding a fork, or the meaning of words or signs; in such situations there is no tension between individual and collective rationality, and thus no room for personal moral norms. It also differs from social norms, which result from "the joint presence of a conditional preference for conformity and the belief that other people will conform as well as approve of conformity" (Bicchieri, Muldoon, & Sontuoso (2023), p.7). Indeed, our model includes neither a pure preference for conformity, nor second-order beliefs about others' approval.

We examine several decision-making situations, varying the individuals' access to information about the others' actions and the marginal impact. We begin by showing that in any decision configuration, each individual's best response, or preferred action, can be simply described by a threshold value, such that she contributes if and only if the share of the others who contribute exceeds it. Depending on the individual's preference type and the belief she holds about the marginal impact, she is either a committed contributor (the threshold equals 0), a committed non-contributor (the threshold exceeds 1), or a conditional contributor (the threshold is between 0 and 1). We show that committed contributors must have a sufficiently strong Kantian concern to overcome aversion towards being behind materially: committed contributors are willing to make a material sacrifice, even if others are not, in order to follow their personal norm. Whether driven or not by some Kantian moral concerns as well, conditional contributors contribute only if sufficiently many others do so: *ceteris paribus*, an increase in an individual's Kantian moral concern and/or aversion towards being ahead materially reduce the threshold whereas aversion towards being behind materially raises it. Finally, committed non-contributors have weak enough Kantian concerns and a weak enough aversion towards being ahead materially. In sum, our model endogenizes the distribution of the individuals' threshold values, simply based on the distribution of their preferences and their beliefs about the marginal impact; this distribution will be seen to play the same role as in the large literature building on the threshold model of collective behavior (Schelling (1978), Granovetter (1978)).

As a benchmark we first characterize Nash equilibria of the one-shot game in which all players select their actions simultaneously and have complete information about the others' preference types as well as the marginal impact. By definition, any Nash equilibrium is self-sustaining: individuals have correct beliefs about the others' actions and nobody wishes to deviate. Multiple equilibria are common. If there are only conditional contributors, the two extreme outcomes (no individual contributes and all individuals contribute) are both in the set of Nash equilibria, which may also contain intermediate outcomes (some individuals contribute). When information about the marginal impact is public, there would be agreement concerning the right thing to do among those individuals who have some Kantian concerns. Hence, the behavioral norm does not necessarily coincide with the personal moral norms: there may be full lack of contributions combined with a full shared understanding that contributing is the right thing to do.

The analysis is then devoted to models with time, assuming that each individual is myopic: his or her beliefs about what the others will do in any given time period are fully determined by what they did in the preceding period.[4] We begin by examining a situation where all the individuals have correct beliefs about the marginal impact and about all the others' actions, and where initially the marginal impact is so low that the moral norm is to not contribute, and the descriptive norm is zero contributions. An exogenous shock, e.g., a technological innovation, increases the marginal impact enough for the moral norm to switch to contributing. While this reduces each individual's threshold, we show that a behavioral change occurs over time if and only if there are some committed contributors under the new moral norm. Such committed contributors all switch to contributing as soon as their personal moral norm has changed; this may then trigger others to switch as well, etc, a process which ends in finite time. The new behavioral norm may or not correspond to full contributions; this depends on the distribution of preferences.

We then consider a setting where individuals are not necessarily correctly informed about the marginal impact. Specifically, all the individuals *falsely* believe that the marginal impact is so low

[4] This is in line with many literature contributions, e.g., Granovetter (1978); Kandori et al. (1993); Young (1993); Brock & Durlauf (2001); Blume & Durlauf (2003); Gavrilets (2021); Gavrilets et al. (2024).

that their personal moral norm is to not contribute, and the descriptive norm is zero contributions. Some individuals then become informed, for example because they read about relevant scientific evidence. This reduces the threshold of those who are informed, and for any behavioral change to occur there must be some individuals who are both informed and committed contributors under the new moral norm. Such leaders, or instigators (Granovetter (1978)), switch to contributing as soon as their personal norm has changed, and this may trigger other informed individuals and/or uninformed individuals to switch as well. Interestingly, uninformed individuals must attach a low enough weight to their Kantian concern to make the switch, since they believe that the right thing to do is to not contribute.

In the next section we describe the model, show how social-Kantian preferences determine the distribution of thresholds, and characterize the set of Nash equilibria in the static game. In Section 3 we analyze the behavioral dynamics in the full network, before proceeding to analysis of other network structures in Section 4. We discuss our findings in light of the literature on social nudges in Section 5, before concluding.

2 Setup and benchmark

2.1 The material game

A population with a finite number N of individuals faces a collective action problem, formalized as a public goods game. Each individual $i \in I = \{1, 2, ..., N\}$ either contributes ($x_i = 1$) or not ($x_i = 0$). The net material benefit for i from own action and others' actions, described by the $(N-1)$-dimensional vector $\boldsymbol{x}_{-i}$, is

$$\pi_i(x_i, \boldsymbol{x}_{-i}) = \left(x_i + \sum_{j \neq i} x_j \right) B - x_i c, \tag{1}$$

where $B > 0$ is the marginal benefit to self of any individual contribution and c is the cost of contributing. Unless stated otherwise, it is collectively rational but individually irrational to contribute in the material game, that is:

$$NB > c > B. \tag{2}$$

Letting s_i denote the share of individuals other than i who contribute,

$$s_i = \frac{\sum_{j \neq i} x_j}{N-1}, \tag{3}$$

the expression in (1) can also be written

$$\pi(x_i, s_i) = s_i(N-1)B + x_i(B-c). \tag{4}$$

This completes the formalization of the*material game* $G = (N, B, c)$.

2.2 Preferences

Preferences, and thus game payoffs, may differ from material payoffs. Moreover, beliefs about the share of others who contribute, and about the marginal benefit B, may be incorrect. Letting these beliefs be denoted $\hat{s}_i$ and $\hat{B}_i$, respectively, we posit that the following utility function describes i's preferences (we will define y below):

$$\begin{aligned} u_i(x_i, \hat{s}_i) \;\; = \;\; & \hat{s}_i(N-1)\hat{B}_i + x_i(\hat{B}_i - c) - \;\; \gamma_i \left(Ny\hat{B}_i - yc - Nx_i\hat{B}_i + x_i c \right) \\ & -\alpha_i c \cdot \max\{0, x_i - \hat{s}_i\} - \beta_i c \cdot \max\{0, \hat{s}_i - x_i\}. \end{aligned} \tag{5}$$

The first two terms in (5) represent the material payoff that i anticipates. The fourth and fifth terms capture the individual's attitude towards being materially behind respectively ahead of the others, in line with formalizations elsewhere of attitudes towards inequity (Fehr & Schmidt (1999)); the public good being non-rival, only the difference between own and others' average contribution cost matters. A strictly positive α_i (resp. β_i) means that i dislikes making a larger (resp. smaller) material sacrifice than the others for the public good. We assume that $\alpha_i + \beta_i \geq 0$: while individuals may like being ahead ($\beta_i < 0$) or behind ($\alpha_i < 0$), we restrict attention to mild versions of such attitudes.[5] Our focus on the comparison between own and others' material payoffs differs from the literature on norms, which has tended to concentrate on a conformity desire, which with our notation is often formalized as a loss $-(s_i - x_i)^2$, unrelated to the cost c; see, e.g., Blume & Durlauf (2003), Blume et al. (2015), Gavrilets (2021), and Arduini et al. (2022).

The third term in (5) measures Kantian moral concerns. It is proportional to the difference between two hypothetical material payoffs. Only the second one ($Nx_i\hat{B}_i - x_i c$) depends on the individual's decision x_i: it is the material payoff the individual believes she would obtain if –

[5] Estimates of these preference parameters in the experimental literature suggest that most individuals are either indifferent or dislike being behind ($\alpha_i \geq 0$), while the attitude towards being ahead is more heterogeneous: some individuals exhibit spite towards others also when ahead ($\beta_i < 0$) whereas others are altruistic ($\beta_i \geq 0$). See, e.g., Bruhin et al. (2018) and van Leeuwen & Alger (2024).

hypothetically – all the other individuals were to use the same strategy that s/he is using, x_i. A positive γ_i thus captures a concern for the material payoff she would obtain if her action was universalized, a concern found to have evolutionary foundations (Alger & Weibull (2013); Alger et al. (2020).

The first hypothetical material payoff in the third term $(Ny\hat{B}_i - yc)$ is the one the individual believes she would obtain if all the other individuals were to use the strategy y, her *personal moral norm*, taken to be the action that i believes would maximize total welfare, if selected by everybody:

$$y = \arg\max_{z\in\{0,1\}} N\left(N\hat{B}_i - c\right)z. \tag{6}$$

Note that the term $Ny\hat{B}_i - yc$ in (5) is a constant. We have included it for ease of comparison with the literature on social norms, where a standard assumption is that individuals incur a psychological cost from deviating from their personal norm, i.e., the strategy that they hold as being "the right thing to do". Our specification differs in two ways. First, it endogenizes the personal norm, as being the action i's believes would maximize her material payoff if it was universalized (see (6)). Second, the disutility from deviating from the personal norm is also based on a universalization argument, which induces the individual to evaluate each action in the light of the loss in own material payoff that would follow it everybody were to select action x_i instead of y. A key difference with the standard approach is thus that the personal norm and the utility from deviating from it both depend on the specifics of the material game (the parameters N and c) and the individual's beliefs $\hat{B}_i$ about the marginal benefit B. By contrast, in the literature the cost of deviating from the personal norm is formalized by way of a loss $-(y - x_i)^2$, which is unrelated to $\hat{B}$, c, and N; see, e.g, D'Adda et al. (2020) and Gavrilets (2021).

Henceforth, we refer to $\theta_i = (\alpha_i, \beta_i, \gamma_i)$ as individual i's *preference type*, and let $\Theta = \{\theta_1, \theta_2, ..., \theta_N\}$ denote the *preference profile* in the population. While in the general analysis we make no specific assumptions about this distribution (except that $\alpha_i + \beta_i \geq 0$ for all $i \in I$), throughout the paper we will illustrate the theoretical results using the estimates of the preference types $\theta_i = (\alpha_i, \beta_i, \gamma)$ for 95 of the subjects who participated in the experimental study of van Leeuwen & Alger (2024) (among the 112 subjects included in their main analysis we exclude those whose estimates violate our assumption $\alpha_i + \beta_i \geq 0$). These estimates are included in Table 1 in the Appendix.[6]

[6] The preference specification of van Leeuwen & Alger (2024) is indeed equivalent to ours when the reciprocity parameters (δ_i and γ_i in their equation (1)) are set to 0, and their expression is divided through by $1 - \kappa_i$. In other

2.3 Best responses

We first establish some results on best responses. Throughout we impose the tie-breaking assumption that i contributes if indifferent. Given the beliefs $\hat{s}_i$ and $\hat{B}_i$, i thus contributes if and only if

$$\hat{B}_i - c + \gamma_i(N\hat{B}_i - c) - \alpha_i c(1 - \hat{s}_i) \geq -\beta_i c \hat{s}_i. \tag{7}$$

If $\alpha_i + \beta_i = 0$ (e.g., if i is purely Kantian with $\alpha_i = \beta_i = 0$), this condition is independent of $\hat{s}_i$ and boils down to

$$\gamma_i \geq \frac{(1+\alpha_i)c - \hat{B}_i}{N\hat{B}_i - c}. \tag{8}$$

In words, individuals who attach the same weight to the others' material payoffs whether ahead or behind ($\alpha_i = -\beta_i$), including those who are purely Kantian ($\alpha_i = \beta_i = 0$), and who derive utility from following their personal norm, contribute regardless of how many others do so, as long as their Kantian concern is sufficiently pronounced.

For any individual for whom $\alpha_i + \beta_i > 0$, we rewrite (7) as a condition on the minimum share of others contributing for i to contribute it as well:

$$\hat{s}_i \geq \frac{(1+\alpha_i)c - \hat{B}_i - \gamma_i(N\hat{B}_i - c)}{(\alpha_i + \beta_i)c}. \tag{9}$$

Defining

$$\tilde{s}(\theta_i, \hat{B}_i) = \begin{cases} 0 \text{ if } \frac{(1+\alpha_i)c - \hat{B}_i - \gamma_i(N\hat{B}_i - c)}{(\alpha_i+\beta_i)c} \leq 0 \\ \frac{(1+\alpha_i)c - \hat{B}_i - \gamma_i(N\hat{B}_i - c)}{(\alpha_i+\beta_i)c} \text{ otherwise,} \end{cases} \tag{10}$$

we have thus established:

Proposition 1. *The preference type θ_i of each individual i together with her belief $\hat{B}_i$ about the marginal benefit B uniquely determines a threshold $\tilde{s}(\theta_i, \hat{B}_i)$, such that i contributes if and only if she believes that the share of others who contribute exceeds it, $\hat{s}_i \geq \tilde{s}(\theta_i, \hat{B}_i)$. This threshold is decreasing in the perceived benefit $\hat{B}_i$ and increasing in the cost c.*

Our model thus determines endogenously each individual's threshold, taken to be exogenous in the literature based on Granovetter's (1978) model.[7] Moreover, our model establishes a link between the thresholds and the specifics of the material game, or, more precisely, the individuals'

words, our α_i corresponds to their $\frac{\alpha_i}{1-\kappa_i}$, our β_i to their $\frac{\beta_i}{1-\kappa_i}$, and our γ_i to their $\frac{\kappa_i}{1-\kappa_i}$. The estimates we use are the ones reported in Section IV.A of van Leeuwen & Alger (2024).

[7] For recent contributions, see, e.g., Centola et al. (2018) and Wiedermann et al. (2020).

perceived material benefits and costs. This implies that policy interventions aiming at correcting these beliefs may affect behavior by way of altering the individual thresholds and thus the threshold distribution.

By examining the individual thresholds, we see that there may be those whose Kantian concerns are strong enough for them to contribute regardless of others' actions; indeed, $\tilde{s}(\theta_i, \hat{B}_i) = 0$ if

$$\gamma_i \geq \frac{(1+\alpha_i)c - \hat{B}_i}{N\hat{B}_i - c}. \tag{11}$$

There may also be those who do not contribute regardless of others' actions. Indeed, $\tilde{s}(\theta_i, \hat{B}_i) > 1$ if i does not suffer too much from being materially ahead ($\beta_i < 1 - \hat{B}_i/c$) and has a small enough Kantian concern:

$$\gamma_i < \frac{(1-\beta_i)c - \hat{B}_i}{N\hat{B}_i - c}, \tag{12}$$

where the right-hand side is positive if and only if $\beta_i < 1 - \hat{B}_i/c$. Defining

$$\tilde{\gamma}(z, \hat{B}_i) = \frac{(1+z)c - \hat{B}_i}{N\hat{B}_i - c} \tag{13}$$

and

$$\tilde{\beta}(\hat{B}_i) = 1 - \hat{B}_i/c, \tag{14}$$

we obtain the following result (the comparative statics results of point 3 are straightforward):

Proposition 2. *Given the cost (c), her preferences θ_i and her beliefs about the marginal benefit ($\hat{B}_i$), individual i is:*

1. ***a committed contributor**, for whom contributing is a dominant strategy, if $\gamma_i \geq \tilde{\gamma}(\alpha_i, \hat{B}_i)$;*
2. ***a committed non-contributor**, for whom not contributing is a dominant strategy, if $\beta_i < \tilde{\beta}(\hat{B}_i)$ and $\gamma_i < \tilde{\gamma}(-\beta_i, \hat{B}_i)$;*
3. ***a conditional contributor**, who contributes if and only if they believe that the share of others is at least $\tilde{s}(\theta_i, \hat{B}_i)$, if $\tilde{\gamma}(-\beta_i, \hat{B}_i) \leq \gamma_i < \tilde{\gamma}(\alpha_i, \hat{B}_i)$; the threshold value $\tilde{s}(\theta_i, \hat{B}_i)$ is decreasing in γ_i, β_i, and $\hat{B}_i$, and increasing in α_i and c.*

This proposition shows that our model offers a preference-based explanation for why some individuals may be insensitive to the share of others who contribute, while others are. Those committed to not contributing do not suffer much from being ahead of others or from deviating from the personal norm; those committed to contributing have a sufficiently pronounced Kantian concern. By contrast, a conditional contributor contributes if and only if they perceive that sufficiently many others do so. The threshold depends on the individual's preferences in expected manners: *ceteris paribus*, a higher Kantian concern and a more pronounced aheadness aversion reduces it, while a more pronounced behindness aversion raises it.

The proposition further shows the role that beliefs about the marginal material benefit play: *ceteris paribus*, an increase in $\hat{B}_i$ lowers the individual's threshold $\tilde{s}(\theta_i, \hat{B}_i)$: such an increase can thus potentially transform a conditional contributor to a committed contributor. This points to a second channel through which policy interventions aiming at correcting these beliefs may affect behavior. There is also a third channel, namely, the own material benefit term in (5). We will refer back to these three channels in Section 4, which discusses some field experiments in the light of our model.

Letting $\hat{\mathcal{B}} = (\hat{B}_1, \hat{B}_2, ..., \hat{B}_N)$ denote the profile of beliefs about the marginal impact B, we define the set of committed contributors, $\mathcal{C}(I, \Theta, \hat{\mathcal{B}})$, and $\mathcal{N}(I, \Theta, \hat{\mathcal{B}})$ that of committed non-contributors:

$$\mathcal{C}(I, \Theta, \hat{\mathcal{B}}) = \{i \in I \,|\, \gamma_i \geq \tilde{\gamma}(\alpha_i, \hat{B}_i)\} \tag{15}$$

$$\mathcal{N}(I, \Theta, \hat{\mathcal{B}}) = \{i \in I \,|\, \beta_i < \hat{\beta}(\hat{B}_i) \text{ and } \gamma_i < \tilde{\gamma}(-\beta_i, \hat{B}_i)\}. \tag{16}$$

When considering settings where beliefs about the marginal impact are correct, i.e., when $\hat{B}_i = B$ for all i, we will omit the argument $\hat{\mathcal{B}}$.

Assuming correct beliefs about B, Figure 1 shows, for $c = 1$ and four different values of B, the distributions of the threshold values $\tilde{s}(\theta_i, B)$ for $N = 95$ and the preference profile in Table 1. Threshold values $\tilde{s}(\theta_i, B)$ below or equal to 0 correspond to the committed contributors, those strictly above 1 to the committed non-contributors, and those between 0 and 1 to the conditional contributors. The threshold values of the conditional contributors are shown using five intervals. As expected, the number of committed contributors is increasing in B while that of committed non-contributors is decreasing. It turns out that given this preference profile the total number of conditional contributors is quite small.

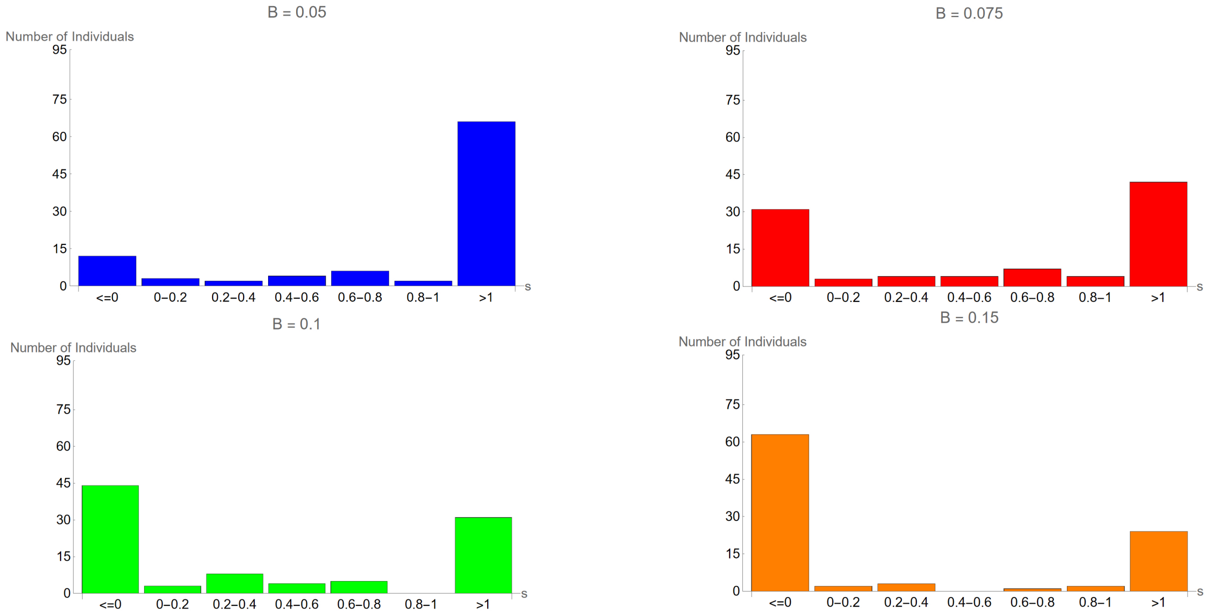


Figure 1: Histograms showing, for $c = 1$ and four different values of B, the number of individuals with threshold values $\tilde{s}(\theta_i, B)$ falling into the seven bins shown on the horizontal axis. $N = 95$ and the preference profile is in Table 1.

2.4 Nash equilibria in the benchmark game

As a benchmark, we consider the game $\Gamma = \langle G, \Theta \rangle$ in which individuals select their strategies simultaneously and under complete information about the material game $G = (N, B, c)$ and the preference profile Θ. The objective is to characterize the set of Nash equilibria of this game. We will describe a Nash equilibrium by referring to the number n^* of individuals who contribute at this equilibrium.

Define the step function $m : \{0, 1/(N-1), 2/(N-1), ..., 1\} \to \{0, 1, ..., N\}$ by

$$m(s) = \#\mathcal{C}(I, \Theta) + \#\{i \in I \setminus \mathcal{C}(I, \Theta) \,|\, \tilde{s}(\theta_i) \leq s\}. \tag{17}$$

This is the total number of individuals who contribute if the number of other individuals who do so is $s(N-1) \in \{0, 1, ..., N-1\}$: they are the committed contributors and the conditional contributors whose threshold value is met. Clearly, a necessary and sufficient condition for exactly $n = (N-1)s + 1$ individuals to contribute is that, given s, the total number of individuals who prefer to contribute equals n:

$$m(s) = (N-1)s + 1. \tag{18}$$

In other words, at a Nash equilibrium empirical expectations must be self-fulfilling.

Proposition 3. *The game $\Gamma = \langle G, \Theta \rangle$ admits at least one Nash equilibrium. Moreover, the set of equilibria is such that the number of contributors is bounded below by $\#\mathcal{C}(I, \Theta)$, and it includes an equilibrium at which:*

1. *$n^* = 0$ if, and only if, $\mathcal{C}(I, \Theta) = \emptyset$, and*

2. *$n^* = N$ if, and only if, $\mathcal{N}(I, \Theta) = \emptyset$.*

Proof. We derive a necessary and sufficient condition for there to exist an equilibrium with some positive number of individuals $n^* \in \{1, 2, ..., N\}$ who contribute. Let $s^* = (n^* - 1)/(N - 1)$. A necessary and sufficient condition for exactly $n^* = s^*(N - 1) + 1$ individuals to contribute is that, given s^*, the total number of individuals who prefer to contribute equal n^*:

$$m(s^*) = (N - 1)s^* + 1. \tag{19}$$

There are thus two cases to consider:

1. $m(0) = 0$: this is a sufficient and necessary condition for $n^* = 0$ to be a Nash equilibrium;

2. $m(0) > 0$: since n is defined for each $s \in \{0, 1/(N-1), 2/(N-1), ..., 1\}$, is weakly increasing in s, and is bounded above by N, while the number $(N - 1)s + 1$ is strictly increasing in s and reaches the value N for $s = 1$, there must exist some $s \in \{0, 1/(N-1), 2/(N-1), ..., 1\}$ such that $m(s) = (N - 1)s + 1$; moreover, $n^* = 0$ is then not a Nash equilibrium.

Whether $m(0) = 0$ or not, there thus exists a Nash equilibrium. Furthermore, $\mathcal{C}(I, \Theta) \neq \emptyset$ is clearly a sufficient condition for $m(0) > 0$. It is also necessary, since not contributing is a best response to $s = 0$ for any conditional contributor i. Finally, $\mathcal{N}(I, \Theta) = \emptyset$ is obviously a necessary condition for $n^* = N$ to be a Nash equilibrium; it is also sufficient, since by definition, any individual who does not belong to $\mathcal{N}(I, \Theta)$ contributes if all the others do so. □

In words, while an equilibrium always exists, a full lack of contributions ($n^* = 0$) is not always an equilibrium, and neither is generalized contributions ($n^* = N$). For the former to exist there must not be any committed contributors; this is also a sufficient condition, because any conditional contributor prefers not to contribute if nobody else does. For the latter to exist, there must not be any committed non-contributors; this is also a sufficient condition, because any conditional contributor contributes if everybody else does.

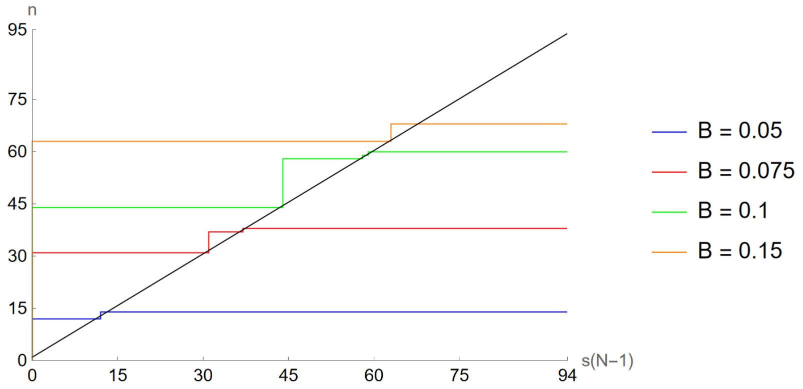
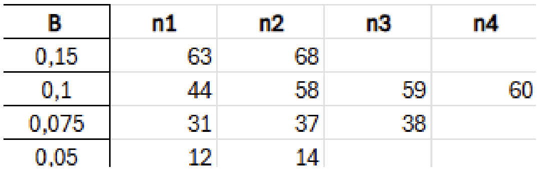


Figure 2: The step function m for $B = 0.05$ (bottom), $B = 0.075$ (second from bottom), $B = 0.1$ (third from bottom), and $B = 0.15$ (top). The straight is $n = (N-1)s + 1$. $N = 95$ and the preference profile is in Table 1.

A necessary and sufficient condition for exactly $n^* = s^*(N-1)+1 \geq 1$ individuals to contribute is that, given s^*, the total number of individuals who prefer to contribute equal n^*:

$$m(s^*) = (N-1)s^* + 1. \tag{20}$$

For $N = 95$ and the preference profile in Table 1, Figure 2 shows, for $c = 1$ and four different values of B, the step function m. Any intersection between the step function and the upward-sloping line, which corresponds to $n = (N-1)s + 1$, is a Nash equilibrium. Figure 3 shows the number of contributions n^* at all the Nash equilibria thus identified.



B	n1	n2	n3	n4
0,15	63	68		
0,1	44	58	59	60
0,075	31	37	38	
0,05	12	14		

Figure 3: Number of contributions (n^*) at all the Nash equilibria for $B = 0.05$ (bottom), $B = 0.075$ (second from bottom), $B = 0.1$ (third from bottom), and $B = 0.15$ (top). $N = 95$ and the preference profile is in Table 1.

In the next proposition we characterize the set of Nash equilibria in a population with only two preference types, $\theta_A = (\alpha_A, \beta_A, \gamma_A)$ and $\theta_B = (\alpha_B, \beta_B, \gamma_B)$, with I_A and I_B denoting the set of type-A and type-B individuals, respectively. In what follows, we identify Nash equilibria by the number of players contributing in them.

Proposition 4. *Consider a population where $0 < N_A < N$ individuals have preference type $\theta_A = (\alpha_A, \beta_A, \gamma_A)$ and $N_B = N - N_A$ individuals have preference type $\theta_B = (\alpha_B, \beta_B, \gamma_B)$.*

1. *If $\mathcal{C}(I, \Theta) = \mathcal{N}(I, \Theta) = \emptyset$, then $n^* = 0$ and $n^* = N$ are Nash equilibria; moreover, $n^* = N_X$, $X = A, B$, is a Nash equilibrium if, and only if, $\tilde{s}(\theta_X) \leq (N_X - 1)/(N - 1)$ and $\tilde{s}(\theta_Y) > N_X/(N - 1)$, where $Y \neq X$.*

2. *If $\mathcal{C}(I, \Theta) \neq \emptyset$ and $\mathcal{N}(I, \Theta) = \emptyset$, then $n^* = N$ is a Nash equilibrium; moreover, $n^* = N_X$, $X = A, B$ is a Nash equilibrium if, and only if $\mathcal{C}(I, \Theta) = I_X$ and $\tilde{s}(\theta_Y) > N_X/(N-1)$, where $Y \neq X$.*

3. *If $\mathcal{C}(I, \Theta) = \emptyset$ and $\mathcal{N}(I, \Theta) \neq \emptyset$, then $n^* = 0$ is a Nash equilibrium; moreover, $n^* = N_X$, $X = A, B$ is a Nash equilibrium if, and only if $\mathcal{N}(I, \Theta) = I_Y$ and $\tilde{s}(\theta_X) \leq (N_X - 1)/(N - 1)$, where $Y \neq X$.*

4. *If $\mathcal{C}(I, \Theta) = I_X$ and $\mathcal{N}(I, \Theta) = I_Y$, then $n^* = N_X$ is the unique Nash equilibrium.*

A formal proof is omitted as the statements about $n^* = 0$ and $n^* = N$ follow directly from Proposition 3, and the other statements follow from condition (9) in a straightforward way. For example, if $\mathcal{C}(I, \Theta) = \mathcal{N}(I, \Theta) = \emptyset$, $n^* = N_A$ is a Nash equilibrium if and only if $\tilde{s}(\theta_A) \leq (N_A - 1)/(N - 1)$ so that each type-A individual contributes as long as all the other type-A individuals do so, and and $\tilde{s}(\theta_B) > N_B/(N - 1)$, so that type-B individuals do not contribute even if all the type-A individuals contribute.

We illustrate the implications of qualitatively different preference types, by describing the set of equilibria in four examples. Define the threshold values

$$\hat{\gamma}(\alpha_i, \beta_i, s) = \frac{c - B + \alpha_i c - (\alpha_i + \beta_i) cs}{NB - c} \tag{21}$$

and

$$\hat{\beta}(\alpha_i, \gamma_i, s) = \frac{c - B + \alpha_i (1 - s) c - \gamma_i (NB - c)}{cs}, \tag{22}$$

and let type θ_k, $k = A, B$, be:

- *Homo oeconomicus* if $\alpha_k = \beta_k = \gamma_k = 0$;
- only inequity-averse if $\alpha_k > 0$, $\beta_k > 0$, $\gamma_k = 0$;
- only Kantian if $\alpha_k = 0$, $\beta_k = 0$, $\gamma_k > 0$;
- Kantian and inequity-averse if $\alpha_k > 0$, $\beta_k > 0$, $\gamma_k > 0$.

Example 1 (A is only inequity-averse, B is *Homo oeconomicus*)**.** *The B-type is a committed non-contributor, while the A-type contributes if and only if all the other A-types contribute and their aheadness aversion is sufficiently pronounced, $\beta_A \geq \hat{\beta}(\alpha_A, 0, (N_A - 1)/(N - 1))$. There is thus a unique Nash equilibrium with $n^* = 0$ if $\beta_A < \hat{\beta}(\alpha_A, 0, (N_A - 1)/(N - 1))$, and two Nash equilibria, with $n^* = 0$ and $n^* = N_A$, respectively, otherwise.*

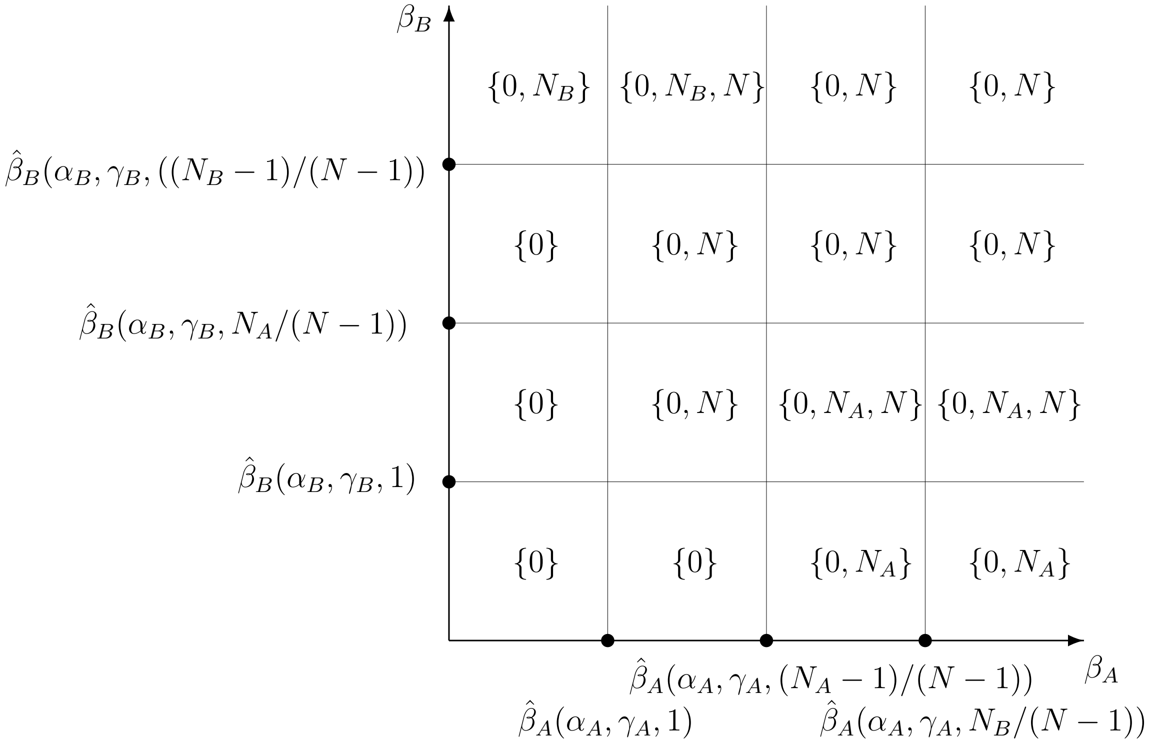
In this example,

Example 2 (A is only Kantian, B is *Homo oeconomicus*)**.** *The B-type is a committed non-contributor, while the A-type is a committed contributor if $\gamma_A \geq \hat{\gamma}(0, 0, 0)$ and a committed non-contributor if $\gamma_A < \hat{\gamma}(0, 0, 0)$. Hence, there is a unique Nash equilibrium, at which $n^* = N_A$, if $\gamma_A \geq \hat{\gamma}(0, 0, 0)$, and a unique Nash equilibrium, at which $n^* = 0$, if $\gamma_A < \hat{\gamma}(0, 0, 0)$.*

Example 3 (A is only Kantian (and strongly so), B is only inequity-averse)**.** *Suppose the A-type has $\gamma_A > \hat{\gamma}(0, 0, 0)$ so that it is a committed contributor. Turning to the B-type, there are three cases. First, if $\beta_B < \hat{\beta}(\alpha_B, 0, 1)$, its aheadness aversion is so low that it prefers not to contribute even if everyone else contributes. There is then a unique equilibrium, at which $n^* = N_A$. Second, if $\beta_B \geq \hat{\beta}(\alpha_B, 0, N_A/(N - 1))$, its aheadness aversion is so pronounced that it contributes as long as all the A-types do so. There is then a unique equilibrium, at which $n^* = N$. Finally, if $\hat{\beta}(\alpha_B, 0, 1) \leq \beta_B < \hat{\beta}(\alpha_B, 0, N_A/(N - 1))$, the B-type contributes if everyone else does, but not if only the A-types do so. There are then two equilibria, one at which $n^* = N_A$ and one at which $n^* = N$.*

Example 4 (A is Kantian and inequity-averse type, B is only inequity-averse)**.** *The B-type is the same as in Example 3. The A-type is a committed contributor if $\gamma_A \geq \hat{\gamma}(\alpha_A, \beta_A, 0)$; applying the same logic as in the preceding example, we conclude that there is then a unique equilibrium, at which $n^* = N_A$, if $\beta_B < \hat{\beta}(\alpha_B, 0, 1)$, a unique equilibrium, at which $n^* = N$, if $\beta_B \geq \hat{\beta}(\alpha_B, 0, N_A/(N-1))$, and two equilibria, one at which $n^* = N_A$ and one at which $n^* = N$, if $\hat{\beta}(\alpha_B, 0, 1) \leq \beta_B < \hat{\beta}(\alpha_B, 0, N_A/(N - 1))$. By contrast, if $\gamma_A < \hat{\gamma}(\alpha_A, \beta_A, 0)$, no type is a committed contributor. Figure 4 shows how the set of equilibria then depends on the values of β_A and β_B.*

These examples show that equilibrium multiplicity obtains for qualitatively different preference distributions. As stated in the proposition, preference distributions where at least some individuals have strong enough Kantian concerns eliminate the sustainability of the most socially suboptimal


β_B
$\hat{\beta}_B(\alpha_B, \gamma_B, ((N_B - 1)/(N - 1))$
$\hat{\beta}_B(\alpha_B, \gamma_B, N_A/(N - 1))$
$\hat{\beta}_B(\alpha_B, \gamma_B, 1)$
$\hat{\beta}_A(\alpha_A, \gamma_A, (N_A - 1)/(N - 1))$
β_A
$\hat{\beta}_A(\alpha_A, \gamma_A, 1)$
$\hat{\beta}_A(\alpha_A, \gamma_A, N_B/(N - 1))$

$\{0, N_B\}$	$\{0, N_B, N\}$	$\{0, N\}$	$\{0, N\}$
$\{0\}$	$\{0, N\}$	$\{0, N\}$	$\{0, N\}$
$\{0\}$	$\{0, N\}$	$\{0, N_A, N\}$	$\{0, N_A, N\}$
$\{0\}$	$\{0\}$	$\{0, N_A\}$	$\{0, N_A\}$

Figure 4: The set of Nash equilibrium contributions n^* for different combinations of β_A (on the horizontal axis) and β_B (on the vertical axis); $N_B < N_A - 1$.

outcome $x^* = 0$, where nobody contributes (see Examples 2, 3, and 4). Furthermore, the combination of a sufficiently Kantian type with a sufficiently aheadness-averse type makes the socially optimal outcome $x^* = N$ sustainable as a Nash equilibrium (see Examples 3 and 4). However, Kantian concerns are not necessary for the existence of equilibria with contributions, as long as some individuals are sufficiently averse to being materially better off than others (see Examples 1 and 4).

Figure 4 further shows how the set of the possible number of equilibrium contributions depends on the aversion towards being ahead, with two preference types, and assuming that $N_B < N_A - 1$. The horizontal axis shows β_A and the vertical axis β_B. On each axis there are three threshold values. We here assume that there are no committed contributors: $n^* = 0$ is a Nash equilibrium for any (β_A, β_B). We then see that $n^* = N$ is a Nash equilibrium if and only if the A-type and the B-type are sufficiently aheadness averse to be conditional contributors (i.e., $\beta_A \geq \hat{\beta}_A(\alpha_A, \gamma_A, 1)$ and

$\beta_B \geq \hat{\beta}_B(\alpha_B, \gamma_B, 1)$). Equilibria where only the A-type (respectively only the B-type) contributes arise if the β_A is large enough and β_B is small enough (respectively β_B is large enough and β_A is small enough).

Two further remarks on Nash equilibria are called for. First, depending on the preference profile, coincidence between the behavioral norm (i.e., the most common behavior at equilibrium) and the personal moral norm (which given that B is common knowledge is the same for all individuals with some Kantian moral concerns) may clearly fail to arise. This is the case either when the number of committed non-contributors constitutes the majority of the population, or the number of committed contributors is too small to induce sufficiently many conditional contributors to contribute.

Second, consider the following two settings, not covered above: (a) $B > c$, so that contributing is both individually and collectively rational; (b) $c > NB$, so that contributing is neither individually nor collectively rational. In both cases, the preference distribution may be such that all individuals are conditional contributors, in which case both $n^* = 0$ and $n^* = N$ are Nash equilibria.

3 Dynamics

We now consider the spread of pro-social behavior over time. Suppose that the simultaneous-move interaction among N individuals described above takes place at each point in (discrete) time t, and that in each period individuals best-respond to the actions undertaken in the last period, which are public information. In other words, individuals are fully myopic and form their beliefs based solely on past observed behavior. The empirical expectation of individual i at time period t is thus

$$\hat{s}_{i,t} = \frac{\sum_{j \neq i} x_{j,t-1}}{N-1}. \tag{23}$$

Our focus is on scenarios where not contributing is initially a dominant strategy for all individuals, and where some exogenous shock occurs. We examine whether this shock triggers behavioral changes. We analyze two settings. In the first, beliefs about the marginal impact are correct, and there is a change of the marginal impact at some point in time. In the second, beliefs marginal impact may be incorrect, and they may also differ between individuals.

3.1 A publicly observable increase in B

Initially, the publicly observable marginal benefit (B_0) is such that contributing is neither individually nor collectively rational, that is, $c > NB_0 \geq B_0$. Suppose further that initially no individual contributes. This is compatible with the posited preferences and the belief $\hat{s}_{i,t} = 0$, since for any $(\alpha_i, \beta_i, \gamma_i) \in \mathbb{R}^2_+$ individual i strictly prefers to not contribute, since (recall (7))

$$B_0 - c - \alpha_i c + \gamma_i (NB_0 - c) < 0. \tag{24}$$

In other words, initially there are no committed contributors, $\mathcal{C}_0(I, \Theta) = \emptyset$. At some point in time, say $t = 0$, some change (e.g., in technology or the environment) occurs which increases the benefit to some $B > B_0$ sufficiently large to make contributing collectively rational, $NB - c > 0$, while still being individually irrational, $c > B$. This change being publicly observable, the personal norm switches from not contributing to contributing (recall (6)). Given our assumptions on how beliefs are formed, at time $t = 1$ each individual i holds the belief $\hat{s}_{i,1} = 0$, and thus switches to contributing only if their Kantian concern exceeds the threshold value $\tilde{\gamma}(\alpha_i, \beta_i, 0)$ (recall (21)).

Proposition 5. *Suppose that at time $t = 0$ the marginal benefit from contributing increases sufficiently to render contributing collectively rational, $NB - c > 0$, while still being individually irrational, $c > B$. Suppose that prior to the change no individual contributed, and that individuals are myopic. Then, the increase in B generates some behavioral change at time $t = 1$ if and only if it entails some committed contributions, i.e., $\mathcal{C}(I, \Theta) \neq \emptyset$.*

Henceforth, let $\mathcal{C}(I, \Theta) \neq \emptyset$, so that the number of individuals who contribute at time $t = 1$, denoted n_1, is:

$$n_1 = \#\mathcal{C}(I, \Theta), \tag{25}$$

leading any individual i who did *not* contribute at $t = 1$ to hold the belief for $t = 2$ equal to

$$\hat{s}_{i,2} = \frac{n_1}{N-1}. \tag{26}$$

At time $t = 2$ any such individual thus contributes if and only if

$$B - c + \gamma_i(NB - c) - \alpha_i c \left(1 - \frac{n_1}{N-1}\right) \geq -\beta_i \frac{c\,n_1}{N-1}, \tag{27}$$

or (recalling (21))

$$\gamma_i \geq \hat{\gamma}\left(\alpha_i, \beta_i, \frac{n_1}{N-1}\right). \tag{28}$$

Since $\hat{\gamma}(\alpha_i, \beta_i, n_1/(N-1)) < \hat{\gamma}(\alpha_i, \beta_i, 0)$, the total number of informed individuals who contribute at time $t = 2$ is:

$$n_2 = \#\{i \in I \mid \gamma_i \geq \hat{\gamma}(\alpha_i, \beta_i, n_1/(N-1))\}. \tag{29}$$

More generally, at any point in time, any individual i who did *not* contribute in period $t-1$ holds the belief for period t equal to

$$\hat{s}_{i,t} = \frac{n_{t-1}}{N-1}, \tag{30}$$

and contributes at time t if and only if

$$\gamma_i \geq \hat{\gamma}(\alpha_i, \beta_i, n_{t-1}/(N-1)), \tag{31}$$

so that the total number of informed individuals who contribute at time t is:

$$n_t = \#\{i \in I \mid \gamma_i \geq \hat{\gamma}(\alpha_i, \beta_i, n_{t-1}/(N-1))\}. \tag{32}$$

Ceteris paribus, $\hat{\gamma}$ is decreasing in n_{t-1}. It follows that any individual who switched from not contributing to contributing at some point in time, will never switch back to not contributing. Since the population is finite, eventually, players will stop updating their actions, yielding the following result:

Proposition 6. *Suppose that at time $t = 0$ the marginal benefit from contributing increases sufficiently to render contributing collectively rational, $NB - c > 0$, while still being individually irrational, $c > B$. Suppose that prior to the change no individual contributed, and that individuals are myopic. For any preference profile Θ, there exists a finite $\hat{t} \geq 1$ such that no further behavioral changes occur after time period $\hat{t}$.*

Figure 5 shows, in two ways, what these dynamics would have been if $N = 95$ and the preference distribution had been the one in Table 1, for four different values of the marginal impact B; starting from the bottom line and moving upwards, $B = 0.05$, $B = 0.075$, $B = 0.1$, and $B = 0.15$. The left panel shows, for each value of B, the sequence n_1, s_1, n_2, s_2, etc, while the right panel shows the ensuing sequence of contributions as a function of time.

We see that all the dynamics are stabilized after between two and four periods. In the right panel the dashed horizontal lines indicate the maximum number of contributions, equal to the total number of individuals minus the committed non-contributors. While none of the dynamics reaches this maximum, a higher B induces a higher share of the potential contributors to contribute.

Moreover, comparing the left panel of Figure 5 with Figure 2, we see that the dynamics all reach the number of contributions corresponding to the Nash equilibrium with the highest n^*. This is driven by the combination of a sizable share of committed contributors and a relatively even distribution of threshold values $\tilde{s}(\cdot)$ among the conditional contributors (recall Figure 1).

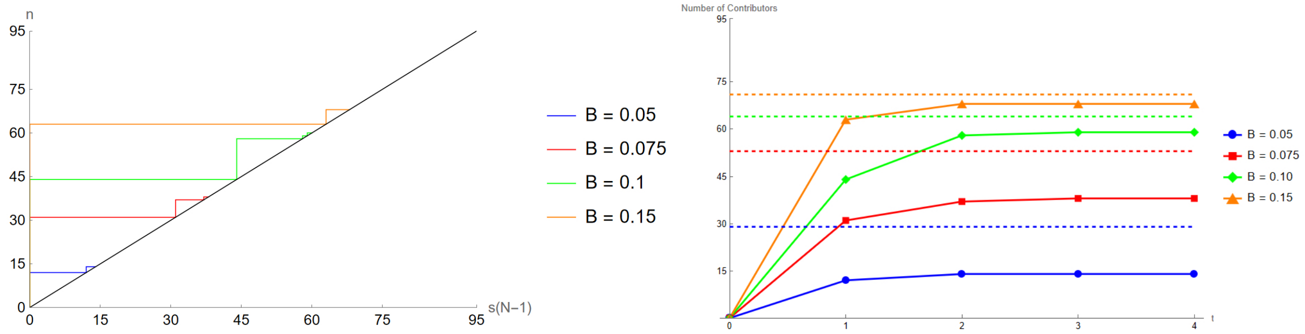


Figure 5: Contribution dynamics for $B = 0.05$ (bottom), $B = 0.075$ (second from bottom), $B = 0.1$ (third from bottom), and $B = 0.15$ (top). Left panel: the straight line is the 45-degree line. Right panel: the dashed lines show the total number of potential contributors (committed and conditional contributors).

3.2 Heterogenous beliefs about the marginal benefit B

We now discard the assumption that individuals have correct beliefs about the marginal benefit. Suppose that initially all individuals *falsely believe* that the marginal benefit is so low that contributing is collectively irrational, that is, the initial common belief $\hat{B}_i = \hat{B}$ for all i is such that $N\hat{B} < c$. Initially, they thus all falsely believe that the "right thing to do" is to not contribute. Suppose further that initially no individual contributes. This is compatible with the posited preferences and the belief $s_{i,t} = 0$, since for any $(\alpha_i, \beta_i, \gamma_i) \in \mathbb{R}^2_+$ individual i strictly prefers to not contribute, since (recall (7))

$$\hat{B} - c - \alpha_i c + \gamma_i \left(N\hat{B} - c \right) < 0. \qquad (33)$$

At some point in time, say $t = 0$, the beliefs of a set $J \subseteq I$ of individuals are corrected, for example thanks to a governmental information campaign or press coverage of a scientific publication. Assuming that every informed individual $j \in J$ instantaneously switches their belief to $\hat{B}_j = B$, their personal norm switches to "contribute". Hence, j contributes at time $t = 1$ if and only if $\gamma_j \geq \hat{\gamma}(\alpha_j, \beta_j, 0)$. We first focus on individuals who at time $t = 1$ is (a) correctly informed that contributing is the right thing to do, and (b) willing to contribute even if nobody else does because of a strong enough Kantian motivation. Formally, this defines the set of *leaders* as

$$\mathcal{L}(J) = \{j \in J \,|\, \gamma_j \geq \hat{\gamma}(\alpha_j, \beta_j, 0)\}. \tag{34}$$

Turning now to the uninformed individuals, their personal norm is still “do not contribute”. Some uninformed individuals may nonetheless start contributing if sufficiently many informed individuals have done so, and if they are sufficiently averse to being better off than others materially. Specifically, this happens if for some uninformed individual k and some empirical expectation $s_{k,t}$:

$$\hat{B} - c - \alpha_k c(1 - s_{k,t}) + \gamma_k \left(N\hat{B} - c\right) \geq -\beta_k c s_{k,t}, \tag{35}$$

or

$$\gamma_k \leq \frac{\hat{B} - c - \alpha_k c + (\alpha_k + \beta_k) c s_{k,t}}{c - N\hat{B}} = \hat{\gamma}_0(\alpha_k, \beta_k, s_{k,t}), \tag{36}$$

where an index 0 has been added to the threshold value to differentiate it from the one corresponding to the correct belief B about the marginal benefit. Since uninformed individuals believe that not contributing is the right thing to do, they need to attach a *small* enough importance to the Kantian moral concern in order to start contributing.

Hence, the total number of individuals who contribute at time t is:

$$n_t = \#\{j \in J \mid \gamma_j \geq \hat{\gamma}(\alpha_j, \beta_j, n_{t-1}/(N-1))\} \; + \; \#\{k \notin J \mid \gamma_k \leq \hat{\gamma}_0(\alpha_k, \beta_k, n_{t-1}/(N-1))\}. \tag{37}$$

Ceteris paribus, $\hat{\gamma}$ is decreasing and $\hat{\gamma}_0$ is increasing in n_{t-1}. It follows that any individual who switched from not contributing to contributing at some point in time, will never switch back to not contributing. Moreover, at any point in time there may be both informed and uninformed individuals who contribute. Hence:

Proposition 7. *Suppose that at times $t < 0$ all individuals incorrectly believe that not contributing is the right thing to do. At $t = 0$ a set $J \subseteq I$ of individuals obtain the correct information about B. Suppose that prior to the change no individual contributed, and that individuals are myopic. Then:*

1. *There exists a finite $\hat{t} \geq 1$ such that no further behavioral changes occur after time period $\hat{t}$.*
2. *The information dissemination has some effect on behavior ($n_{\hat{t}} \geq 1$) if, and only if, there is at least one leader ($\mathcal{L}(J) \neq \emptyset$).*

3. *Depending on the preference profile Θ and the information dissemination J, if at $\hat{t}$ there are still some individuals who do not contribute, these may be uninformed and/or informed.*

The two main take-aways from this analysis are (a) that the correct information must reach at least one individual with a sufficiently pronounced Kantian moral concern for it to have any effect on behavior, and (b) that information is neither sufficient nor necessary for individuals other than these leaders to switch from not contributing to contributing.

4 Discussion

As evidence against the idea that the purely materially self-interested *Homo oeconomicus* is found in every human being accumulates, economists are becoming increasingly interested in non-monetary policy instruments (see, e.g, Thaler & Sunstein (2008), Bowles (2016)). This has motivated field experiments studying the effectiveness of a variety of informational interventions. Areas of application in line with our model include energy and water conservation, recycling, public transportation usage, and tax compliance; these interventions indeed concern individually costly but collectively beneficial behaviors. Our model is relevant for two types of interventions: social comparisons, and correction of (first-order) beliefs.[8] Here we discuss the findings of some such field experiments in the light of our model. Our aim is *not* to provide an overview, but rather to make a few observations to highlight how our model might be useful for such experiments.

Social comparison interventions provide individuals with feedback on behavior: own and that of people in some reference group. Our model predicts that if prior to the intervention individuals underestimate (respectively overestimate) others' efforts, the intervention should enhance (respectively reduce) their willingness to provide effort. If, moreover, an individual's pre-intervention effort level is positively correlated with these beliefs, one should expect to see mixed results, whereby those whose high (respectively low) pre-intervention efforts reduce (respectively increase) their efforts. Such a "boomerang effect" has been observed in some field experiments on electricity consumption (e.g., Schultz et al. (2007)). However, other studies document an asymmetric response: increased effort (on average) among those with pre-intervention below-average efforts and either small positive or nil effects on those with pre-intervention above-average efforts; see, e.g.,

[8] Second-order beliefs about others' normative views being absent from our model, we do not discuss studies that document effects of correcting such beliefs (e.g., Bursztyn et al. (2020)).

Allcott (2011) for a study on electricity consumption, and Ferraro & Price (2013) for a study on water consumption. Such an asymmetric response can be explained by models relying solely on conformity bias if first-order beliefs about others' behaviors are asymmetric. By contrast, in our model such asymmetries can arise due to preferences, e.g., if there are more committed contributors than committed non-contributors. This is in line with the finding of Schultz et al. (2016) that individuals with stronger personal norms (measured by way of questions such as "I feel a personal obligation to save as much water as possible," and "I feel morally obliged to save water, regardless of what others do") were less sensitive to social comparison feedback.

A key difference between our model and models based on pure conformity bias, is that the thresholds that trigger pro-social behavior depend on their perceived benefits as well as their costs (recall Proposition 1). This may help reconcile the positive effects of social comparisons on electricity and water consumption reported above with the nil effects on public transportation usage found by Gravert & Olsson Collentine (2021). While the pro-environmental behaviors in question are all beneficial for one's wallet, reducing one's electricity or water consumption at the margin arguably entails a much smaller reduction in comfort than would a switch from car to public transportation. And while the latter would also induce a larger benefit on the environment, on balance it may be that the thresholds tend to be larger for public transportation usage than for energy and water conservation.

Some field experiments provide participants with factual information about the benefits that the behavior in question entails. This has been found to have positive effects on, e.g., tax compliance (Bott et al. (2020)) and food waste recycling (Linder et al. (2018)). In our model, if such information implies an upward correction of the perceived benefit ($\hat{B}_i$ in the model), it can affect behavior through three channels: first, because it corrects the belief about the own net material benefit ($\hat{B}_i - c$) upwards; second, because the individual's threshold is reduced, as mentioned above; and, third, because it can lead to an updward adjustment of the personal norm (recall equation (6)). By contrast, in models relying solely on conformity bias, only the first effect would be at work. Our model thus suggests that correction of first-order beliefs about costs and benefits, combined with social comparison feedback, should have a greater overall effect than if the two interventions are carried out separately.

5 Concluding remarks

Humans are complex social animals, whose behavior is often influenced by that of others. A large number of factors are potentially at work behind such social influences. We have proposed a model that focuses on three such factors: (a) the individuals' beliefs about the material benefits and costs of behavior; (b) their Kantian moral motivations, which together with the said beliefs determine both their personal moral norms and their willingness to follow this norm even if this entails making a substantial material sacrifice compared to others; and (c) their attitudes towards being materially ahead and behind others. For any given distribution of preferences and beliefs, this model is rich enough to produce endogenously a unique distribution of individual thresholds for collectively desirable behavior. The model produces empirically testable predictions, which arguably are useful for field studies which aim at evaluating the potential power of information-based policy interventions. In particular, our model suggests that interventions which combine the provision of correct information about the material benefits and costs with information about others' behavior, should lead to greater behavioral adjustments than if these interventions are carried out separately.

Needless to say, to capture the full complexity of norms and norm change, second-order beliefs about what others deem appropriate should also be included, since these have been shown to matter in several field experiments (see, e.g., Schultz et al. (2007), Bursztyn et al. (2020)). Conformity bias could also be a factor, although the evidence from field experiments that information about others' behavior matters, could also be explained by attitudes towards being materially ahead and/or behind. Recent work further suggests that group identity may affect norms and norm change (Ehret et al. (2022)). To account for this, we are currently working on enriching the model by embedding individuals in a network structure. Finally, punishment may play a role in shaping norms and norm change, a question that has been the subject of a large number of contributions (see, e.g., Gavrilets & Richerson (2017) and Molho et al. (2024) and references therein).

References

Akerlof, G. A. (1997). Social distance and social decisions. *Econometrica*, *65*(5), 1005–1027.

Akerlof, G. A., & Kranton, R. E. (2000). Economics and Identity. *Quarterly Journal of Economics*, *115*(3), 715–753.

Alger, I., Gavrilets, S., & Durkee, P. (2024). Proximate and ultimate drivers of norms and norm change. *Current Opinion in Psychology*, *60*, 101916.

Alger, I., & Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, *81*(6), 2269–2302.

Alger, I., Weibull, J. W., & Lehmann, L. (2020). Evolution of preferences in structured populations: Genes, guns, and culture. *Journal of Economic Theory*, *185*, 104951.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9), 1082–1095.

Andrighetto, G., Gavrilets, S., Gelfand, M., Mace, R., & Vriens, E. (2024). Social norm change: drivers and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1897), 20230023.

Arduini, T., Bisin, A., Özgür, O., & Patacchini, E. (2022). *Dynamic social interactions and smoking behavior* (Working Paper).

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, *102*(5), 841–877.

Bicchieri, C. (1990). Norms of cooperation. *Ethics*, *100*(4), 838–861.

Bicchieri, C. (2006). *The Grammar of Society.* Cambridge: Cambridge University Press.

Bicchieri, C., Dimant, E., Gelfand, M., & Sonderegger, S. (2023). Social norms and behavior change: The interdisciplinary research frontier. *Journal of Economic Behavior and Organization*, *205*, A4–A7.

Bicchieri, C., Muldoon, R., & Sontuoso, A. (2023). Social norms. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2023 ed.). Metaphysics Research Lab, Stanford University.

Binmore, K. (1998). *Just Playing: Game Theory and the Social Contract Vol. 2.* Cambridge, MA: MIT Press.

Bisin, A., Horst, U., & Özgür, O. (2006). Rational expectations equilibria of economies with local interactions. *Journal of Economic Theory*, *127*(1), 74–116.

Blume, L., Brock, W., Durlauf, S., & Jayaraman, R. (2015, apr). Linear Social Interactions Models. *Journal of Political Economy*, *123*(2), 444–496.

Blume, L., & Durlauf, S. (2003, sep). Equilibrium Concepts for Social Interaction Models. *International Game Theory Review*, *05*(03), 193–209.

Bott, K. M., Cappelen, A. W., Sørensen, E. O., & Tungodden, B. (2020). You've got mail: A randomized field experiment on tax evasion. *Management Science*, *66*(7), 2801-2819.

Bowles, S. (2016). *The Moral Economy – Why Good Incentives Are No Substitute for Good Citizens* . New Haven, CT: Yale University Press.

Brekke, K. A., Kverndokk, S., & Nyborg, K. (2003). An economic model of moral motivation. *Journal of Public Economics*, *87*(9), 1967–1983.

Brock, W. A., & Durlauf, S. N. (2001). Discrete choice with social interactions. *Review of Economic Studies*, *68*(2), 235–260.

Bruhin, A., Fehr, E., & Schunk, D. (2018). The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association*, *17*(4), 1025–1069.

Bursztyn, L., González, A. L., & Yanagizawa-Drott, D. (2020). Misperceived social norms: women working outside the home in saudi arabia. *American Economic Review*, *110*(10), 2997–3029.

Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, *360*(6393), 1116-1119.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, *55*, 591–621.

D'Adda, G., Dufwenberg, M., Passarelli, F., & Tabellini, G. (2020). Social norms with private values: Theory and experiments. *Games and Economic Behavior*, *124*(2020), 288–304.

Ehret, S., Constantino, S. M., Weber, E. U., Efferson, C., & Vogt, S. (2022). Group identities can undermine social tipping after intervention. *Nature Human Behaviour*, *6*(12), 1669–1679.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, *3*(4), 99–117.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*(3), 817–868.

Ferraro, P. J., & Price, M. K. (2013). Using nonpecuniary strategies to influence behavior: evidence from a large-scale field experiment. *Review of Economics and Statistics*, *95*(1), 64–73.

Gavrilets, S. (2020). The dynamics of injunctive social norms. *Evolutionary Human Sciences*, *2*, e60.

Gavrilets, S. (2021). Coevolution of actions , personal norms and beliefs about others in social dilemmas. *Evolutionary Human Sciences*, *3*(e44).

Gavrilets, S., & Richerson, P. J. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences*, *114*(23), 6068–6073.

Gavrilets, S., Tverskoi, D., & Sánchez, A. (2024). Modelling social norms : an integration of the norm-utility approach with beliefs dynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(20230027).

Gelfand, M. J., Gavrilets, S., & Nunn, N. (2024). Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change. *Annual Review of Psychology*, *75*, 341–378.

Glaeser, E., & Scheinkman, J. A. (2000). *Non-market interactions* (Working Paper 8053). NBER.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, *83*(6), 1420-1443.

Gravert, C., & Olsson Collentine, L. (2021). When nudges aren't enough: Norms, incentives and habit formation in public transport usage. *Journal of Economic Behavior & Organization*, *190*, 1–14.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: The emergence of costly punishment. *Science*, *316*(5833), 1905–1907.

Heyes, C. (2024). Rethinking Norm Psychology. *Perspectives on Psychological Science*, *19*(1), 12–38.

Kandori, M., Mailath, G. J., & Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, *61*(1), 29–56.

Kuran, T. (1989). Sparks and prairie fires: A theory of unanticipated political revolution. *Public Choice*, *61*(1), 41–74.

Kuran, T., & Sandholm, W. H. (2008). Cultural integration and its discontents. *Review of Economic Studies*, *75*(1), 201–228.

Lindbeck, A., Nyberg, S., & Weibull, J. W. (1999). Social norms and economic incentives in the welfare state. *Quarterly Journal of Economics*, *114*(1), 1–35.

Linder, N., Lindahl, T., & Borgström, S. (2018). Using Behavioural Insights to Promote Food Waste Recycling in Urban Households — Evidence From a Longitudinal Field Experiment. *Frontiers in Psychology*, *9*(March), 1–13.

López-Pérez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior*, *64*(1), 237–267.

Miettinen, T., Kosfeld, M., Fehr, E., & Weibull, J. W. (2020). Revealed preferences in a sequential prisoners' dilemma: a horse-race between six utility functions. *Journal of Economic Behavior and Organization*, *173*, 1-25.

Molho, C., De Petrillo, F., Garfield, Z. H., & Slewe, S. (2024). Cross-societal variation in norm enforcement systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1897), 17–20.

Nyborg, K. (2018). Social norms and the environment [Journal Article]. *Annual Review of Resource Economics*, *10*, 405-423.

Nyborg, K., & Rege, M. (2003). On social norms: the evolution of considerate smoking behavior. *Journal of Economic Behavior & Organization*, *52*(3), 323-340.

Richerson, P. J., Gavrilets, S., & De Waal, F. B. (2021). Modern theories of human evolution foreshadowed by Darwin's Descent of Man. *Science*, *372*(6544).

Schelling, T. C. (1978). *Micromotives and Macrobehavior.* New York, NY: Norton.

Schultz, P. W., Messina, A., Tronu, G., Limas, E. F., Gupta, R., & Estrada, M. (2016). Personalized normative feedback and the moderating role of personal norms: A field experiment to reduce residential water consumption. *Environment and Behavior*, *48*(5), 686–710.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychological Science*, *18*(5), 429–434.

te Velde, V. L. (2022). Heterogeneous norms: Social image and social pressure when people disagree. *Journal of Economic Behavior & Organization*, *194*, 319–340.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness.* New Haven, CT: Yale University Press.

Ullman-Margarit, E. (1977). *The Emergence of Norms.* Oxford: Oxford University Press.

van Leeuwen, B., & Alger, I. (2024). Estimating Social Preferences and Kantian Morality in Strategic Interactions. *Journal of Political Economy Microeconomics*, *2*(4), 665–706.

Wiedermann, M., Smith, E. K., Heitzig, J., & Donges, J. F. (2020). A network-based microfoundation of Granovetter's threshold model for social tipping. *Scientific Reports*, *10*(1), 11202.

Young, H. P. (1993). The evolution of conventions. *Econometrica*, *61*(1), 57–84.

Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics*, *7*(1), 359–387.

Appendix

Table 1: Estimates of $\theta_i = (\alpha_i, \beta_i, \gamma_i)$

α_i	β_i	γ_i	α_i	β_i	γ_i	α_i	β_i	γ_i	α_i	β_i	γ_i
-0,09	1,05	2,62	0,03	0,48	0,24	0,36	0,02	0,12	-0,02	0,12	0,04
2,21	-1,41	1,08	0,37	-0,08	0,22	0,26	0,16	0,12	0,17	-0,06	0,04
0,44	0,42	0,88	0,19	0,51	0,21	0,15	0,32	0,11	0,07	0,05	0,03
0,46	1,98	0,83	0,36	0,35	0,21	0,09	0,33	0,11	0,09	0,48	0,03
-0,05	0,46	0,56	0,51	-0,01	0,20	0,09	0,19	0,11	0,11	0,14	0,03
-0,09	0,54	0,54	0,49	0,19	0,20	0,14	0,33	0,10	0,18	0,20	0,02
0,17	0,19	0,45	0,35	0,15	0,20	0,23	0,51	0,10	0,05	0,42	0,02
0,94	-0,57	0,44	0,18	-0,13	0,19	0,06	0,48	0,10	-0,05	0,16	0,02
0,52	-0,20	0,41	0,39	-0,02	0,19	0,12	0,33	0,10	0,65	0,22	0,01
0,54	0,39	0,37	0,05	1,03	0,18	0,10	0,26	0,10	0,00	0,15	0,01
0,27	-0,02	0,32	0,09	0,57	0,18	0,10	0,19	0,10	0,00	0,30	0,01
0,25	0,06	0,32	0,10	0,58	0,16	0,30	0,32	0,10	0,04	0,31	0,01
0,53	-0,05	0,31	0,13	0,47	0,15	0,25	0,19	0,09	-0,05	0,56	0,00
0,59	-0,21	0,31	0,32	0,61	0,15	0,46	0,08	0,09	0,06	0,09	0,00
-0,07	0,30	0,29	0,01	0,19	0,15	0,03	0,45	0,08	0,08	0,21	0,00
0,48	-0,35	0,29	0,38	-0,07	0,14	0,18	0,19	0,08	-0,07	0,49	0,00
-0,01	0,59	0,28	0,02	0,39	0,14	0,33	0,56	0,08	-0,08	0,67	0,00
-0,07	0,36	0,27	0,22	0,34	0,13	0,21	0,66	0,07	0,05	0,56	0,00
0,46	-0,36	0,27	0,28	0,03	0,13	0,37	0,13	0,06	0,06	0,65	0,00
0,24	0,43	0,26	0,11	0,22	0,13	0,04	0,36	0,06	0,03	0,80	0,00
0,39	0,38	0,26	0,36	0,09	0,13	-0,02	0,21	0,06	0,22	0,73	0,00
0,07	0,14	0,25	0,13	0,30	0,13	0,15	0,15	0,05	0,31	0,75	0,00
0,35	-0,15	0,24	0,09	0,17	0,12	0,15	0,25	0,05	0,07	0,42	0,00
0,22	-0,21	0,24	0,50	0,90	0,12	0,05	0,22	0,04			

To obtain the values in the table, we use the estimates of the behindness aversion, aheadness aversion, and Kantian concern parameters obtained by van Leeuwen & Alger (2024) (called α_i, β_i, and κ_i in their article) when they assume that subjects are risk neutral and exhibit no reciprocity (i.e., the parameters δ_i and γ_i in the preference specification in their equation (1) are set to 0). For each subject i we then divided these estimates by $(1 - \kappa_i)$ to get estimates corresponding to our parameters α_i, β_i, and γ_i (recall Footnote 2). Moreover, to be in line with our assumptions, we restrict attention to the 95 subjects among the 112 core subjects in van Leeuwen & Alger (2024) for whom $\alpha_i + \beta_i \geq 0$.