WORKING PAPERS

N° 1605

April 2025

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

Ingela Alger

Toulouse School of Economics

Norms and norm change — driven by social-Kantian preferences*

Ingela Alger†

April 3, 2025

Abstract

Norms indicate which behaviors are common and/or considered morally right. This paper analyzes norms and norm change by incorporating two hitherto neglected factors: Kantian moral concerns and attitudes towards making a greater or a smaller material sacrifice than others. In an N-person social dilemma, these preferences determine individuals' personal moral norms and their thresholds for collective behavior (cooperation is conditional on sufficiently many others cooperating). Conditions on preferences and beliefs promoting/hampering changes in the behavioral norm (the modal behavior) are identified. Implications for policy interventions aimed at changing norms are discussed in light of the model.

Keywords: personal moral norms, behavioral norms, beliefs, social-Kantian preferences, social norms

*I acknowledge 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 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

1 Introduction

Recycling, refraining from bribing, picking up one's dog's leavings, respecting the order of arrival in queues, washing one's hands every now and then, stepping outside to smoke at a party. In many countries, most of these behaviors, which generate positive externalities at a cost to their authors, are nowadays considered natural: 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. In addition, costly behaviors that do not generate positive externalities, such as female genital mutilation and cutting, are sometimes sustained as norms (Congdon Fors et al. (2024)). Interest among social scientists for norms and norm change have led to experimental interventions, some of which have triggered significant behavioral changes, for example in energy and water conservation (Schultz et al. (2007), Allcott (2011), Schultz et al. (2016)), in female labor market participation (Bursztyn et al. (2020)), in tax evasion (Bott et al. (2020), Besley et al. (2023)), and in social distancing during the COVID-19 pandemic (Vriens et al. (2024)). This paper makes a contribution to the theoretical literature on social norms, conventions, and other behavioral regularities, by proposing a model that can explain norm variation over time and across space.[1]

The literature on how norms come about and are sustained (or not) has identified several factors – besides own material payoff – that influence individual behavior:[2]

- the *personal (moral) norm*: the action one believes is "the right thing to do";
- the *descriptive norm*: one's beliefs about others' actions (first-order beliefs);
- the *injunctive norm*: beliefs about others' approval or disapproval of one's actions (second-order beliefs).

Common preference-based explanations for *why* these factors would matter include a desire to conform with others' actions, or conformity bias (Glaeser & Scheinkman (2000), Brock & Durlauf (2001), Blume & Durlauf (2003), Bisin et al. (2006), Blume et al.

[1]For early treatments and recent surveys see Ullman-Margarit (1977), Schelling (1978), Granovetter (1978), Elster (1989), Bicchieri (1990, 2006), Ostrom (2000), Cialdini et al. (1991); Cialdini & Goldstein (2004), Young (1993, 2015), Binmore (1998), Lindbeck et al. (1999), Nyborg & Rege (2003), Hedström (2005), Nyborg et al. (2016), Nyborg (2018), La Ferrara (2019), Bicchieri, Muldoon, & Sontuoso (2023), and Gavrilets et al. (2024).

[2]Many experimental studies, both in lab and in the field, document their impact; for example, Bicchieri (2006), Bicchieri & Xiao (2009), Krupka & Weber (2009), Cardenas (2011), Carlsson et al. (2015), Szekely et al. (2021), Bicchieri et al. (2022), Schram et al. (2022), Andrighetto et al. (2024), Dimant et al. (2024)

(2015), Gavrilets (2021), Arduini et al. (2022), Efferson et al. (2024)), a concern for the perceived social appropriateness of one's actions (Krupka & Weber (2013), Gavrilets (2021), te Velde (2022)), and a concern for the self-image and/or the social image that one's actions generate (Bernheim (1994), Brekke et al. (2003), te Velde (2022), Lane et al. (2023), Bénabou & Tirole (in press)). In this paper I study norms and norm change with a preference class, henceforth social-Kantian preferences, which adds two motivations that have been neglected before: attitudes towards being materially ahead and being behind others (Fehr & Schmidt (1999)), and a Kantian moral concern (Alger & Weibull (2013)).[3] It thus introduces the idea that individuals may be sensitive to making a larger or a smaller sacrifice than others, and [4] and it proposes a theory for how individuals form their personal norms. I will show that this preference class is sufficient to generate spontaneous changes in norms, for easily graspable and intuitive reasons. Moreover, it establishes a tighter link between the specifics of the material consequences of actions on the one hand, and the norm dynamics on the other hand, than previous models.

Social-Kantian preferences differ from those considered elsewhere in the literature on norms in three fundamental ways. First, the Kantian moral concern can explain how an individual forms her personal norm: a universalization argument makes her evaluate each course of action in the light of the material payoff she would obtain if – hypothetically – the others were also to select this course of action. Previous models that do include personal norms take them to be exogenous.[5] Second, the Kantian moral concern determines the utility cost from deviating from the personal norm: it is proportional to the discrepancy between the material payoff the individual would obtain if the action dictated by her personal norm was selected by everyone, and the material payoff she would obtain if her actual action was universalized. Third, the attention paid to the difference between own

[3]Kantian ethics have been around for centuries, and also formalized before by economists (e.g., Laffont 1975, Gravel et al. 2000 Roemer 2019). The social-Kantian preference class examined here emerged recently 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 and survey experiments (Capraro & Rand (2018), Miettinen et al. (2020), Levine et al. (2020), van Leeuwen & Alger (2024)). The study by van Leeuwen & Alger (2024) uses a design that enables estimation of the weights on the social and Kantian motivations. I will use these preference parameter estimates to illustrate the theoretical results.

[4]A recent survey-based study in Austria found that the willingness to make sacrifices was the best predictor for three climate-friendly behaviors (Thaller et al. (2020)).

[5]See Elster (1989), Cialdini & Goldstein (2004),Bicchieri (2006)), D'Adda et al. (2020), Gavrilets (2021), and te Velde (2022), as well as Gavrilets et al. (2024) and references therein. In work that endogenizes how individuals form their opinion about "the right thing to do", such as, for example Brekke et al. (2003) (discussed also by Nyborg (2018)) and López-Pérez (2008), norms are not modeled.

and others' material well-being implies that the utility cost from not conforming with others' actions depends on the material cost, an effect which is absent with conformity bias. Psychologically, the driver is also different from conformity bias: it captures the attitudes towards making a greater or a smaller material sacrifice than others.

In sum, with social-Kantian preferences, one can predict both the individual's personal norm and how she evaluates deviations from others' behavior. The claim is *not* that conformity bias, perceived social appropriateness, and image concerns (as well as other factors which are absent from this model, such as identity (Akerlof & Kranton (2000), Kuran & Sandholm (2008)) and punishment (Gintis et al. (2003), Thöni (2014), Gavrilets & Richerson (2017), Gavrilets (2020), Molho et al. (2024)) are irrelevant; rather, the proposed mechanism behind norms and norm change is complementary. A comprehensive theory would include all the motivations, and empirical studies will be needed to establish whether some motivations appear more relevant than others.

In the model N individuals interact. They are all equipped with social-Kantian preferences, though the weights they attach to the Kantian and the social motivations may differ (they may even be nil). Each individual's preference type is given and fixed through time; the evolution of preferences in the population is not studied, only the evolution of behaviors is. The context is a linear public goods game with two actions, one being costlier than the other. I examine both social dilemmas, where the aggregate material net benefit is maximized if everybody selects the costly action (for example, energy and water conservation, vaccinations, social distancing when ill), and social non-dilemmas, where the aggregate material net benefit is maximized if everybody refrains from the costly action (for example, female genital mutilation). Each individual forms a personal norm based on their beliefs about the material payoff consequences of their actions, and chooses the action that maximizes her utility, given her first-order beliefs (about the others' actions). The two main questions are: What is the set of possible behavioral norms in this population, where a behavioral norm is the most commonly observed behavior at equilibrium?[6] Once a behavioral norm is established, which factors can make it

[6]The concept of behavioral norms 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 (Young (1993)); in such situations there is no tension between individual and collective material interests. 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, the proposed model includes neither a pure preference for conformity, nor second-order beliefs about others' approval.

change? And which factors favor an alignment between behavioral norms and personal moral norms?

Several settings are analyzed, varying the individuals' access to information about the others' actions and the marginal benefit. I begin by showing that in any setting, each individual's best response, or preferred action, can be simply described by a threshold value: she selects the costly action if and only if she believes that the share of the others who do so exceeds this threshold. Depending on the individual's preference type and the belief she holds about the marginal benefit, she is either committed to the costly action (the threshold equals 0), committed to the non-costly action (the threshold exceeds 1), or her action is conditional on her first-order beliefs about the others' actions (the threshold is between 0 and 1). The model thus endogenizes the distribution of the individuals' threshold values, simply based on the distribution of their preferences and their beliefs about the marginal benefit. It thus makes a contribution to the large literature building on the threshold model of collective behavior, in which the threshold distribution is exogenously given (Schelling (1978), Granovetter (1978)). In particular, the model offers a preference-based explanation for why individuals either commit to one or the other action, or condition their behavior on others' behavior (for a recent model which assumes such a classification of individuals, see Wiedermann et al. (2020)).[7]

In social dilemmas, committed contributors are willing to make a material sacrifice even if noboby else does, in order to follow their personal norm: they must thus have a sufficiently strong Kantian concern to overcome aversion towards being behind materially. 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 social non-dilemmas, commitment to the costly action requires a strong enough aversion towards being ahead materially, and a weak enough

[7]Some other models do offer preference-based explanations for the individual thresholds. In a coordination game Andreoni, Nikiforakis, and Siegenthaler (2021) assume that each individual is motivated by own material payoff and an idiosyncratic utility term proportional to the change in material payoff that a collective strategy switch would entail. Gavrilets (2020) assumes that individuals suffer a psychological cost of deviating from the exogenously given norm in the form of disapproval by others. In the model of Mittal et al. (2025), individuals differ in their intrinsic preferences over two goods, but they also benefit when their choice is aligned with that of their neighbors on a network.

Kantian concern, since the personal norm prescribes the non-costly action.

As a benchmark, I 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 benefit. By definition, any Nash equilibrium is self-sustaining: individuals have correct first-order beliefs and nobody wishes to deviate. Multiple equilibria can arise. If there are no committed individuals, the two extreme outcomes (no individual selects the costly action and everybody does) are both in the set of Nash equilibria, which may also contain intermediate outcomes. When information about the marginal benefit is public, there is agreement about the personal norm. Hence, the behavioral norm does not necessarily coincide with the personal norms: there may be full lack of contributions combined with a full shared understanding that contributing is the right thing to do, or *vice versa*.

Attention then turns to norm dynamics, the objective being to examine policy-relevant thought experiments. I assume that each individual holds myopic first-order beliefs, fully determined by the others' behavior in the preceding period.[8]

In the first situation I study (a) individuals hold correct beliefs about the marginal benefit and about the others' past actions, (b) initially the marginal benefit is so low that the personal norm prescribes the non-costly action, and (c) nobody selects the costly action. An exogenous shock, e.g., a technological innovation, increases the marginal benefit enough for the personal norm to switch to the costly action. For example, given the amount of energy required to produce them, the first generation solar panels would not have been viable; more recent ones are around five times as efficient, and environmentally viable in many climates. As a result, each individual's threshold decreases. However, I show that for any spontaneous behavioral change to occur there must be some committed contributors under the new personal norm. Such committed contributors are indeed the only ones to switch to contributing, even though they hold the belief that nobody else contributes; this may then trigger some conditional contributors 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.

In the second policy-relevant thought experiment, the setting is a social dilemma, but

[8]This is in line with many models, e.g., Granovetter (1978); Kandori et al. (1993); Young (1993); Brock & Durlauf (2001); Blume & Durlauf (2003); Efferson et al. (2020); Gavrilets (2021); Efferson et al. (2024); Gavrilets et al. (2024).

individuals are not necessarily correctly informed about the marginal benefit. Specifically, all the individuals *falsely* believe that the marginal benefit is so low that their personal norm is to not contribute. Initially, nobody contributes. Some individuals then become informed, for example because they read about relevant scientific evidence, and this reduces their thresholds. For any behavioral change to occur there must be some individuals who are both informed and committed contributors under the new personal 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.

I also discuss a setting where the beliefs about others' past actions are incorrect, and consider the effects of correcting them. The model is shown to generate ambiguous effects, depending on whether beliefs over- or underestimate others' contributions. Crucially, this depends on the individuals' attitudes towards making a larger or a smaller material sacrifice than others.

The next section describes the setup. Section 3 then derive best responses and shows how social-Kantian preferences determine the distribution of thresholds, and characterizes Nash equilibria of the static game in social dilemmas, while Section 4 does so for social non-dilemmas. In Section 5 I turn to analysis of behavioral dynamics, and discuss some recent research on field interventions in the light of these findings in Section 6, before concluding in Section 7.

2 Setup and benchmark

2.1 The material game

A finite number N of individuals interact in a game, where each individual $i \in I = \{1, 2, ..., N\}$ either undertakes a costly action ($x_i = 1$) 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 $c > 0$ is the cost, and $B \geq 0$ is the benefit that i obtains for each costly action undertaken among the N individuals. Letting s_i denote the share of individuals other than i who contribute,

$$s_i = \frac{\sum_{j \neq i} x_j}{N-1}, \tag{2}$$

the expression in (1) can also be written

$$\pi(x_i, s_i) = s_i(N-1)B + x_i(B-c). \tag{3}$$

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

I will analyze both *social dilemmas*, in which it is materially collectively rational but individually irrational to undertake the costly action,

$$NB > c > B, \tag{4}$$

and *social non-dilemmas*, in which the cost is so large that it is materially both collectively and individually irrational to undertake the costly action,

$$c > NB > B. \tag{5}$$

2.2 Preferences

Preferences, and thus game payoffs, may differ from material payoffs. Moreover, beliefs about the share of others who undertake the costly action as well as 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_i below):

$$\begin{aligned} u_i(x_i, \hat{s}_i, y_i) \;=\; & \hat{s}_i(N-1)\hat{B}_i + x_i(\hat{B}_i - c) - \gamma_i(y_i - x_i)(N\hat{B}_i - c) \\ & -\alpha_i c \cdot \max\{0, x_i - \hat{s}_i\} - \beta_i c \cdot \max\{0, \hat{s}_i - x_i\}. \end{aligned} \tag{6}$$

The first two terms represent the material payoff that i anticipates. The fourth and fifth terms capture material inequity aversion (Fehr & Schmidt (1999)). To see this, note that because the public good (or bad) is non-rival the difference between (what i perceives to be) own material payoff and others' average material payoff reduces to the

difference between own and others' average cost. A strictly positive α_i means that i dislikes obtaining a smaller material payoff than others, and thus captures her sacrifice aversion. A strictly positive β_i means that i dislikes obtaining a larger material payoff than the others, and thus captures her solidarity with them. I assume $\alpha_i + \beta_i \geq 0$: individuals may like being ahead ($\beta_i < 0$) or behind ($\alpha_i < 0$), but attention is restricted to mild such attitudes. I also assume $\beta_i < 1$, ruling out the possibility of attaching a greater weight to the others' average material payoff than to own material payoff.[9] The 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 (using my notation) is often formalized as a utility loss $-(s_i - x_i)^2$ from deviating from the others' average action (see, e.g., Blume & Durlauf (2003), Blume et al. (2015), Gavrilets (2021), and Arduini et al. (2022)). Unlike in my model, this utility loss is unrelated to the cost c.

The third term in (6) measures Kantian moral concerns. It is proportional to the difference between two hypothetical material payoffs. One of them ($x_i(N\hat{B}_i - c)$) depends on i's decision x_i: this is the material payoff she believes she would obtain if – hypothetically – all the others were to use the same action 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 (Alger & Weibull (2013)), and it is assumed that $\gamma_i \geq 0$ for all $i \in I$.

The other hypothetical material payoff in the third term ($y_i(N\hat{B}_i - c)$) is the one i believes she would obtain if all the others were to use action y_i, which is her *personal (moral) norm*, defined as the action that i believes would maximize her material payoff, if it was selected by everybody:

$$y_i = \arg \max_{z \in \{0,1\}} z(N\hat{B}_i - c). \tag{7}$$

While the term $y_i(N\hat{B}_i - c)$ in (6) is a constant, with no implication for i's decision, its inclusion eases the comparison with the literature on social norms. A standard assumption therein is that individuals incur a psychological cost from deviating from their personal norm, which, as here, is the action they hold as "the right thing to do". The specification I propose differs from the standard approach in two ways. First, it endogenizes

[9] 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).

the personal norm, as being the action i believes would maximize her material payoff if it was universalized (see (7)), while in the literature it is exogenously given. Second, the disutility from deviating from the personal norm is also based on a universalization argument: it induces the individual to evaluate each action in the light of the loss in own material payoff that would follow if everybody were to select x_i rather than y_i. 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 as a loss $-(y_i - x_i)^2$, which is unrelated to $\hat{B}$, c, and N; see, e.g, D'Adda et al. (2020) and Gavrilets (2021).

In sum, the utility function in (6) together with the universalization reasoning used to determine the personal norm in (7), makes an individual trade off own material benefit, the utility loss from deviating from the personal norm, and the utility loss (or gain) from making a smaller or a larger material sacrifice than others.

Henceforth, $\theta_i = (\alpha_i, \beta_i, \gamma_i)$ will be referred to as individual i's *preference type*, and $\Theta = \{\theta_1, \theta_2, ..., \theta_N\}$ will denote the *preference profile* in the population. The general analysis makes no specific assumptions about this distribution, except that $\alpha_i + \beta_i \geq 0$ and $\gamma_i \geq 0$ for all $i \in I$. Throughout the paper the theoretical results will be illustrated 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 the assumption $\alpha_i + \beta_i \geq 0$). These estimates are included in Table 1 in the Appendix.[10]

3 Social dilemmas

In a social dilemma all individuals believe that, from a material perspective, it is collectively rational but individually irrational to undertake the costly action:

$$N\hat{B}_i > c > \hat{B}_i \quad \forall i \in I. \tag{8}$$

[10]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 words, the α_i in this model corresponds to their $\frac{\alpha_i}{1-\kappa_i}$, the β_i to their $\frac{\beta_i}{1-\kappa_i}$, and the γ_i to their $\frac{\kappa_i}{1-\kappa_i}$. The estimates I use are the ones corresponding to this utility function, reported in Section IV.A of van Leeuwen & Alger (2024).

For this class of games, i *contributes* (towards the public good) if $x_i = 1$. In this setting, the personal norm is "contribute": $y_i = 1$ for all $i \in I$ (see (7)).

Before turning to analysis of equilibria, results on best responses are established. This will lead to the characterization of individual "thresholds for collective behavior".

3.1 Best responses

Throughout I 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 the utility from contributing is at least as high as the utility from not contributing:

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

If $\alpha_i + \beta_i = 0$, this condition is independent of $\hat{s}_i$ and boils down to

$$\gamma_i \geq \frac{(1+\alpha_i)c - \hat{B}_i}{N\hat{B}_i - c} \equiv \tilde{\gamma}(\alpha_i, \hat{B}_i). \tag{10}$$

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 $\alpha_i + \beta_i > 0$, rewrite (9) as a condition on the minimum share of others contributing for i to contribute it as well,

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

where the right-hand side is negative if $\gamma_i \geq \tilde{\gamma}(\alpha_i, \hat{B}_i)$. Defining

$$\tilde{s}(\theta_i, \hat{B}_i) = \begin{cases} 0 \text{ if } \gamma_i \geq \frac{(1+\alpha_i)c - \hat{B}_i}{N\hat{B}_i - c} \\ \frac{(1+\alpha_i)c - \hat{B}_i - \gamma_i(N\hat{B}_i - c)}{(\alpha_i + \beta_i)c} \text{ otherwise,} \end{cases} \tag{12}$$

the following result has thus been established.

Proposition 1. *Consider a social dilemma. An individual's preference type θ_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, $\hat{s}_i \geq \tilde{s}(\theta_i, \hat{B}_i)$.

In sum, the model determines endogenously each individual's threshold for collective behavior, taken to be exogenous in the literature based on Granovetter's (1978) model.[11] Moreover, for any given preference profile Θ, it establishes a link between the thresholds and the specifics of the material game, or, more precisely, the perceived material benefits and costs. As will be seen in Section 5, this implies that policy interventions aiming at correcting these beliefs may affect behavior by altering the individual thresholds.

Closer examination of the thresholds reveals that there may be individuals who do not contribute regardless of others' actions. Indeed, $\tilde{s}(\theta_i, \hat{B}_i) > 1$ if i's solidarity with others is weak enough,

$$\beta_i < \frac{c - \hat{B}_i - \gamma_i(N\hat{B} - c)}{c} \equiv \tilde{\beta}(\gamma_i, \hat{B}_i). \tag{13}$$

The following result thus obtains (the comparative statics results of point 3 are straightforward):

Proposition 2. *Consider a social dilemma. 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* $\beta_i < \tilde{\beta}(\gamma_i, \hat{B}_i)$*;*

3. ***a conditional contributor****, who contributes if and only if she believes that the share of other contributors is at least* $\tilde{s}(\theta_i, \hat{B}_i)$*, if* $\gamma_i < \tilde{\gamma}(\alpha_i, \hat{B}_i)$ *and* $\beta_i \geq \tilde{\beta}(\gamma_i, \hat{B}_i)$*; the threshold value* $\tilde{s}(\theta_i, \hat{B}_i)$ *is decreasing in* γ_i *and* β_i*, and increasing in* α_i*; it is decreasing in* $\hat{B}_i$ *and* N*, and increasing in* c*.*

The model offers a preference-based explanation for why some individuals may be sensitive to the share of others who contribute, while others are not. A commitment to contribute requires a sufficiently pronounced Kantian concern (γ_i). Those committed

[11] For recent contributions, see, e.g., Centola et al. (2018), Wiedermann et al. (2020), and Andreoni et al. (2021).

to not contributing have a weak enough Kantian concern and a weak enough solidarity towards the others (β_i). Conditional contributors have a weak enough Kantian concern, and possibly also some sacrifice aversion (α_i), to require some others to also make a material sacrifice before doing so, but enough solidarity with the others to contribute if sufficiently many others do so. An individual's threshold depends on 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.

Proposition 2 further highlights a key novelty of the model: the thresholds depend on the specifics of the material game. Thus, any conditional contributor's threshold is increasing in the cost c and decreasing in the perceived benefit $\hat{B}_i$ as well as in the number of individuals N. The effects of c and $\hat{B}_i$ are explained both by the weight attached to own material payoff and to the personal norm: any decrease in the net benefit $\hat{B}_i - c$ reduces the willingness to contribute. The number of individuals N matters because a larger N enhances the utility cost from deviating from the personal norm, thus enhancing the individual's willingness to contribute. The cost parameter further matters for the "comparison with the Joneses" term: an increase in c means a higher utility loss from making a larger sacrifice than the others, and this further reduces the individual's willingness to contribute.

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

$$\mathcal{C}(I, \Theta, \hat{\mathcal{B}}) = \{i \in I \,|\, \gamma_i \geq \tilde{\gamma}(\alpha_i, \hat{B}_i)\} \tag{14}$$

$$\mathcal{N}(I, \Theta, \hat{\mathcal{B}}) = \{i \in I \,|\, \beta_i < \tilde{\beta}(\gamma_i, \hat{B}_i)\}. \tag{15}$$

The argument $\hat{\mathcal{B}}$ will be omitted when considering settings where beliefs about the marginal benefit are correct, i.e., when $\hat{B}_i = B$ for all i.

Noticing that both the threshold value for γ_i above which i is a committed contributor ($\tilde{\gamma}(\alpha_i, \hat{B}_i)$) and the threshold value for β_i below which i is a committed non-contributor ($\tilde{\beta}(\gamma_i, \hat{B}_i)$) are decreasing in $\hat{B}_i$, the following result immediately obtains.

Proposition 3. *Consider a social dilemma. For a given preference profile Θ_i and a given*

cost (c), and some common belief $\hat{B}_i = \hat{B}$ for all i about the marginal benefit B:

1. *the number of committed contributors $\#\mathcal{C}(I, \Theta, \hat{\mathcal{B}})$ is weakly increasing in $\hat{B}$;*

2. *the number of committed non-contributors $\#\mathcal{N}(I, \Theta, \hat{\mathcal{B}})$ is weakly decreasing in $\hat{B}$.*

This result is illustrated in Figure 1, which shows, for $c = 1$ and four different values of $\hat{B}$, the distributions of the threshold values $\tilde{s}(\theta_i, \hat{B})$ for $N = 95$ and the preference profile in Table 1. Threshold values $\tilde{s}(\theta_i, \hat{B})$ below or equal to 0 correspond to the committed

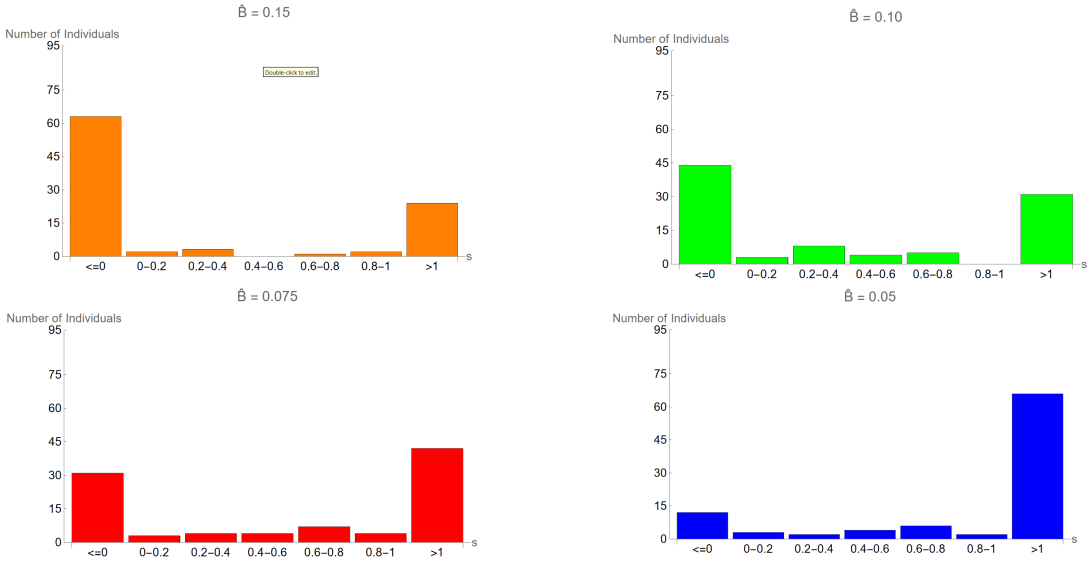

Figure 1: Histograms showing, for $c = 1$ and four different values of $\hat{B}$, the number of individuals with threshold values $\tilde{s}(\theta_i, \hat{B})$ falling into the seven bins shown on the horizontal axis. $N = 95$ and the preference profile is in Table 1.

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. The number of committed contributors increases in $\hat{B}$ while that of committed non-contributors decreases. The figure further shows that with this preference profile the total number of conditional contributors is quite small.

3.2 Nash equilibria in the benchmark game

Here I characterize Nash equilibria of the game $\Gamma = \langle G, \Theta \rangle$ in which individuals select their actions simultaneously and under complete information about the material game $G = (N, B, c)$ and the preference profile Θ. I will describe a Nash equilibrium by referring to the number n^* of individuals who contribute at this equilibrium.

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

$$m(n) = \begin{cases} \#\mathcal{C}(I, \Theta) \text{ if } n = 0 \\ \#\mathcal{C}(I, \Theta) + \#\{i \in I \setminus \mathcal{C}(I, \Theta) \,|\, \tilde{s}(\theta_i) \leq n/(N-1)\} \text{ otherwise,} \end{cases} \tag{16}$$

This is the number of individuals who prefer to contribute as a function of the number of others who do so: they are the committed contributors only if $n = 0$, and the sum of the committed contributors and the conditional contributors whose threshold is met, otherwise. Clearly, $n^* = 0$ is a Nash equilibrium only if $m(0) = 0$, and $n^* \geq 1$ is a Nash equilibrium only if, given that $n^* - 1$ others contribute, n^* individuals are willing to do so, or

$$m\left(n^* - 1\right) = n^*. \tag{17}$$

However, this is not sufficient: the remaining individuals must also prefer to not contribute, given that $n^* \geq 0$ other individuals do contribute:

$$N - n^* = \#\{i \in I \,|\, \tilde{s}(\theta_i) > n^*/(N-1)\}, \tag{18}$$

or, equivalently,

$$m\left(n^*\right) = n^*. \tag{19}$$

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

Proposition 4. *The game $\Gamma = \langle G, \Theta \rangle$ admits at least one Nash equilibrium. 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. There are two cases to consider:

1. $m(0) = 0$: then $n^* = 0$ is a Nash equilibrium, since $m(0) = 0$ means that $\#\mathcal{C}(I, \Theta) = 0$, which in turn implies that condition (18) is met for $n^* = 0$;
2. $m(0) \geq 1$:

(a) since $m(n)$ is weakly increasing in n and is bounded above by $N - \#\mathcal{N}(I, \Theta)$, there exists at least one $n \in \{1, 2, ..., N - \#\mathcal{N}(I, \Theta)\}$ such that $m(n-1) = n$, thereby satisfying condition (17);

(b) among all the values n satisfying $m(n-1) = n$, denote by $\bar{n}$ the largest one;

(c) if $\bar{n} = N - \#\mathcal{N}(I, \Theta)$, then this is a Nash equilibrium, since condition (18) is then satisfied due to the definition of the set of committed non-contributors $\mathcal{N}(I, \Theta)$;

(d) if $\bar{n} < N - \#\mathcal{N}(I, \Theta)$, condition (18) is then satisfied for $n^* = \bar{n}$: supposing by contradiction that it was not satisfied, there would exist some number $a \geq 1$ of individuals who would prefer to contribute given that $\bar{n}$ others do so, and $m(\bar{n} + a - 1) > \bar{n}$; but since $m(n)$ is weakly increasing in n and is bounded above by $N - \#\mathcal{N}(I, \Theta)$, there must then exist at least one $n \in \{m(\bar{n}+a-1), ..., N - \#\mathcal{N}(I, \Theta)\}$ such that $m(n-1) = n$; but this contradicts the definition of $\bar{n}$.

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 $m(0) \geq 1$. 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.

As an illustration, for $N = 95$ and the preference profile in Table 1, Figure 2 shows, for $c = 1$ and four different values of B, the function $m(n)$, as well as the 45-degree line. A Nash equilibrium n^* is such that (a) $m(n^*) = n^*$ (recall (19)), which in the figure is an intersection between $m(n)$ and the 45-degree line, and (b) $m(n^* - 1) = n^*$ (recall (17)),

which in the figure means that n^* is on a horizontal portion of the step function that crosses the 45-degree line. In this case, there is a unique Nash equilibrium for each value of B: $n^* = 68$ if $B = 0.15$, $n^* = 60$ if $B = 0.1$, $n^* = 38$ if $B = 0.075$, and $n^* = 14$ if $B = 0.05$.

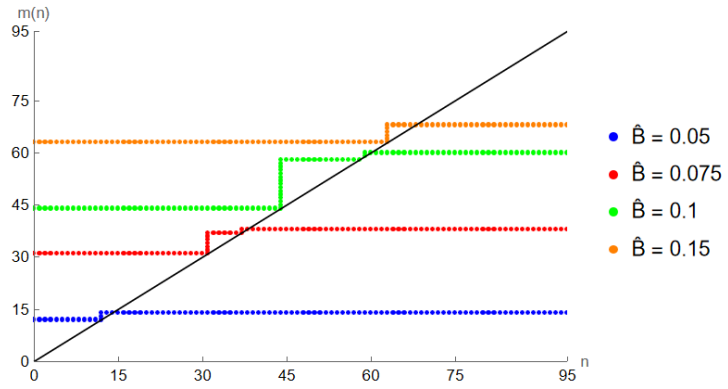

Figure 2: The function m for $B = 0.05$ (bottom), $B = 0.075$ (second from bottom), $B = 0.1$ (third from bottom), and $B = 0.15$ (top). The straight line is the 45-degree line. $N = 95$ and the preference profile is in Table 1.

More generally, however, there may be multiple equilibria. To see this and to better understand how the preference distribution affects the set of Nash equilibria, consider 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 N_A and N_B denoting the number of A-types and B-types, respectively. I illustrate the implications of qualitatively different preference types, by describing the set of equilibria in four examples. Define the threshold values for the Kantian concern and the solidarity parameter

$$\bar{\gamma}(\alpha_i, \beta_i, s, B) = \frac{(1+\alpha_i)c - B - (\alpha_i + \beta_i)cs}{NB - c} \tag{20}$$

and

$$\bar{\beta}(\alpha_i, \gamma_i, s, B) = \frac{c - B - \gamma_i(NB - c) + \alpha_i(1-s)c}{cs}. \tag{21}$$

These are generalizations of the threshold values $\tilde{\gamma}(\alpha_i, \hat{B}_i)$ and $\tilde{\beta}(\gamma_i, \hat{B}_i)$, defined in (10) and (13), to account for values s of the share of others who contribute, different from 0 and 1, respectively. Note that $\bar{\beta}(\alpha_i, \gamma_i, s, B)$ is not defined for $s = 0$, since then the individual cannot be materially better off than the others. 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 $\alpha_k > 0$, $\beta_k > 0$, $\gamma_k > 0$.

In the following examples, for simplicity the argument B has been dropped from the threshold values.

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

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

Example 3 (A is only Kantian (and strongly so), B is only inequity-averse)**.** *Suppose the A-type has $\gamma_A > \bar{\gamma}(0, 0, 0)$ so that it is a committed contributor. Turning to the B-type, there are three cases. First, if $\beta_B < \bar{\beta}(\alpha_B, 0, 1)$, its solidarity towards others is so weak that it is a committed non-contributor. There is then a unique equilibrium, with $n^* = N_A$. Second, if $\beta_B \geq \bar{\beta}(\alpha_B, 0, N_A/(N - 1))$, its solidarity towards others is strong enough for it to be a committed contributor, given that all the A-types contribute. There is then a unique equilibrium, with $n^* = N$. Finally, if $\bar{\beta}(\alpha_B, 0, 1) \leq \beta_B < \bar{\beta}(\alpha_B, 0, N_A/(N - 1))$, the B-type contributes if everyone else does, but not if only the A-types do so. There are then two equilibria, one with $n^* = N_A$ and one with $n^* = N$.*

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

contributors, $n^* = 0$ is a Nash equilibrium for any (β_A, β_B). One then sees that $n^* = N$ is a Nash equilibrium if and only if the A-type and the B-type display a strong enough solidarity with the others (i.e., $\beta_A \geq \bar{\beta}_A(\alpha_A, \gamma_A, 1)$ and $\beta_B \geq \bar{\beta}_B(\alpha_B, \gamma_B, 1)$). 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).

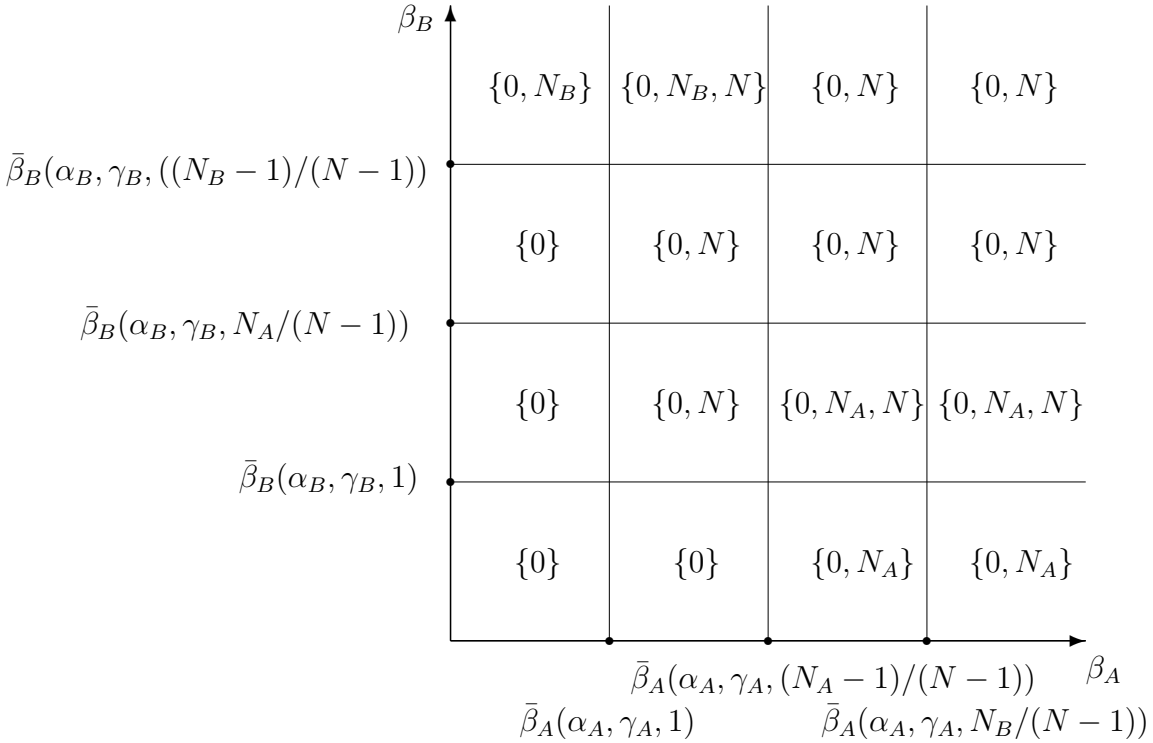

Figure 3: The set of Nash equilibrium contributions n^* if $\gamma_A < \bar{\gamma}(\alpha_A, \beta_A, 0)$ in Example 4, for different combinations of β_A and β_B, and $N_B < N_A - 1$.

These examples show that equilibrium multiplicity obtains for qualitatively different preference distributions. Distributions where at least some individuals have strong enough Kantian concerns eliminate the sustainability of the most socially suboptimal outcome $x^* = 0$, where nobody contributes (see Examples 2, 3, and 4). Furthermore, the combination of a sufficiently Kantian type with a type that exhibits a sufficiently strong solidarity with the others, 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 exhibit a sufficiently strong solidarity with the others (see Examples 1 and 4).

I conclude the analysis of Nash equilibria in the benchmark game by comparing the behavioral norm – that is, the most common behavior at equilibrium – to the personal moral norm "contribute" ($y_i = 1$). The following definitions are adopted:

Definition 1.

- *The modal action at a Nash equilibrium, denoted* $x^* \in \{0, 1\}$*, constitutes the* ***behavioral norm****:* $x^* = 0$ *if* $n^* < N/2$*, and* $x^* = 1$ *if* $n^* \geq N/2$*.*

- *In a population with homogeneity in the personal moral norms,* $y_i = y \in \{0, 1\}$ *for all* $i \in I$*, the behavioral norm is* ***congruent*** *with the personal moral norm if* $x^* = y$*.*

When there are multiple Nash equilibria, these may give rise to the same or to two different behavioral norms. The following result is implied by Proposition 4.

Corollary 1. *In a game* $\Gamma = \langle G, \Theta \rangle$*:*

- *there exists a Nash equilibrium at which the behavioral norm is congruent with the personal moral norm if and only if* $\mathcal{N}(I, \Theta) < N/2$*;*

- *the behavioral norm is congruent with the personal moral norm at all Nash equilibria if and only if* $\mathcal{C}(I, \Theta) \geq N/2$*.*

In other words, if the number of individuals with weak enough Kantian concerns and solidarity with others for them to be committed non-contributors exceeds half of the population, any Nash equilibrium entails non-congruence between the behavioral norm and the personal moral norm. Conversely, if the number of individuals with strong enough Kantian concerns and solidarity with others for them to be committed contributors exceeds half of the population, any Nash equilibrium entails the said congruence.

4 Social non-dilemmas

In a social non-dilemma, all individuals believe that, from a material perspective, it is both collectively and individually irrational to undertake the costly action:

$$c > N\hat{B}_i \geq \hat{B}_i \quad \forall i \in I. \tag{22}$$

The costly action generates a public benefit $N\hat{B}_i \geq 0$ (or at least no harm), but the cost is so large that it is highly inefficient. In this setting, the personal norm is "refrain from the costly action": $y_i = 0$ for all $i \in I$ (see (7)). The analysis proceeds as in the section on social dilemmas, with a focus on the differences with that setting.

4.1 Best responses

Condition (9) is still necessary and sufficient for i to undertake the costly action. By contrast to the social dilemma setting, however, the costly action cannot be a dominant strategy. To see this, consider first the case $\alpha_i + \beta_i = 0$, so that (9) becomes:

$$\gamma_i \leq \frac{(1+\alpha_i)c - \hat{B}_i}{N\hat{B}_i - c}. \tag{23}$$

In other words, i's Kantian concern must be *small* enough for the costly action to be viable. However, this condition is violated, because the right-hand side is strictly negative: indeed, the numerator is negative due to the assumption (22), while the assumption $\beta_i < 1$ implies that the numerator is strictly positive (since $\alpha_i + \beta_i = 0$ then implies that $\alpha_i > -1$).

Next, if $\alpha_i + \beta_i > 0$, i selects the costly action if and only if sufficiently many others do so, like in the social dilemma setting. Indeed, condition (11) still holds. By contrast to the social dilemma setting, however, the right-hand side of this condition is strictly positive (due to assumption (22)). Hence, the threshold value

$$\tilde{s}(\theta_i, \hat{B}_i) = \frac{(1+\alpha_i)c - \hat{B}_i + \gamma_i(c - N\hat{B}_i)}{(\alpha_i + \beta_i)c} \tag{24}$$

is now strictly positive. Intuitively, since the personal norm is to refrain from the costly action, an individual would undertake it only if some others undertake it, in which case the driving force would be a strong enough solidarity towards the others. It is a dominant strategy to not undertake the costly action if, and only if, $\tilde{s}(\theta_i, \hat{B}_i) > 1$, that is, if i does not suffer too much from being materially ahead,

$$\beta_i < \frac{c - \hat{B}_i + \gamma_i(c - N\hat{B}_i)}{c}, \tag{25}$$

that is, if $\beta_i < \tilde{\beta}(\gamma_i, \hat{B}_i)$ (recall (13)). This proves (the comparative statics results of

point 2 are straightforward):

Proposition 5. *Consider a social non-dilemma. Given the cost (c), her preferences θ_i and her beliefs about the marginal benefit ($\hat{B}_i$), for individual i:*

1. *undertaking the costly action cannot be a dominant strategy;*
2. *not undertaking the costly action is a dominant strategy if $\beta_i < \tilde{\beta}(\gamma_i, \hat{B}_i)$;*
3. *it is optimal to undertake the costly action if and only if i believes that the share of others who do so is at least $\tilde{s}(\theta_i, \hat{B}_i)$, as defined in (24), if $\beta \geq \tilde{\beta}(\gamma_i, \hat{B}_i)$; the threshold value $\tilde{s}(\theta_i, \hat{B}_i)$ is decreasing in β_i, $\hat{B}_i$, and N, and increasing in γ_i, α_i, and c.*

By contrast to the social dilemma setting, in social non-dilemmas individuals undertake the costly action only due to their solidarity towards others: their β_i must be high enough *and* sufficiently many others must undertake the costly action. Another difference is that *ceteris paribus* a stronger Kantian concern raises the threshold value, while in the social dilemma it reduces it. Because the "right thing to do" here consists in not undertaking the costly action, a higher γ_i implies that, for any given value of β_i, a larger number of others undertaking the costly action is needed for the solidarity with them to outweigh the utility cost from deviating from the personal norm. In sum, no individual is committed to undertaking the costly action, while the set $\mathcal{N}(I, \Theta, \hat{\mathcal{B}})$, of individuals who are committed to abstaining from the costly action, defined in (15), may be non-empty. The last part of the proposition shows that the thresholds are still increasing in the cost c, and decreasing in the perceived benefit $\hat{B}_i$ and the number of individuals N, as intuition would suggest.

4.2 Nash equilibria in the benchmark game

Consider the game $\Gamma = \langle G, \Theta \rangle$ in which individuals select their strategies simultaneously and under complete information about the preference profile Θ and the material game $G = (N, B, c)$, which satisfies assumption (22) and is thus a social non-dilemma.

The analysis conducted for the social dilemma setting in subsection 3.2 carries over as is. A qualitative difference arises, however, since here there are no individuals who undertake the costly action unconditionally (the set $\mathcal{C}(\cdot)$ is irrelevant). Proposition 4 thus

implies that $n^* = 0$ is a Nash equilibrium for any preference distribution. At this equilibrium, the behavioral norm is congruent with the personal norm (recall Definition 1). However, there may also exist equilibria with non-congruence. In particular, Proposition 4 and Proposition 5 together imply that $n^* = N$ is a Nash equilibrium as long as nobody has strong enough Kantian concerns ($\gamma_i \leq \tilde{\gamma}(-\beta_i, B)$ for all $i \in I$). More generally:

Corollary 2. *Given a social non-dilemma* $\Gamma = \langle G, \Theta \rangle$*:*

- *there exists at least one Nash equilibrium where the behavioral norm is congruent with the personal norm;*
- *the behavioral norm is congruent with the personal norm at all Nash equilibria if and only if* $\mathcal{N}(I, \Theta) \geq N/2$*.*

By contrast to social dilemmas, where "doing the right thing" is costly, here it entails refraining from incurring the cost. This in turn implies that congruence between the behavioral norm and the personal moral norms is more easily achievable than in social dilemmas: in social non-dilemmas there always exists an equilibrium with congruence.

5 Dynamics

The above analysis shows that the preference distribution and the beliefs about the material benefits of the costly action together determine the personal norms and the distribution of thresholds for collective behavior. I will now examine how behavioral norms may change over time, the objective being to evaluate the possible consequences of plausible policy interventions. To this end, 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 taken to be public information. In other words, individuals are fully myopic and form their first-order beliefs based solely on observed past behavior.[12] Individual i thus assumes that the share of individuals who will select the costly action in time period t is

$$\hat{s}_{i,t} = \frac{\sum_{j \neq i} x_{j,t-1}}{N-1}. \tag{26}$$

[12]This assumption is in line with many extant models (Young (1993), Brock & Durlauf (2001), Blume & Durlauf (2003), Acemoglu & Jackson (2015),Besley et al. (2023)), although some models assume forward-looking agents (Bisin et al. (2006)). Yet other models restrict attention to static equilibria (D'Adda et al. (2020), te Velde (2022), Bénabou & Tirole (in press)).

Interactions which from a material standpoint are social dilemmas and social non-dilemmas will be examined in turn, and n_t will denote the number of individuals who select the costly action at time t.

5.1 Social dilemmas

Consider a situation where not contributing is initially a dominant strategy for all individuals, and where some exogenous shock occurs. Does this shock trigger any behavioral changes? We analyze two policy-relevant settings. In the first, beliefs about the benefit are correct, and there is a change in the benefit at some point in time. In the second, beliefs about the benefit may be incorrect, and they may also differ between individuals; the shock consists in correcting some individuals' beliefs.

5.1.1 A publicly observable increase in B

Consider a population whose size N remains fixed through time, and that there is some action whose cost c is also fixed through time. Initially, until some point in time $t = 0$, the publicly observable benefit (B_0) is so low that it is a social non-dilemma. Suppose further that until $t = 0$, in every period no individual undertook the costly action. This is compatible with the posited preferences and the belief $\hat{s}_{i,t} = 0$, because for any $(\alpha_i, \beta_i, \gamma_i)$ individual i prefers $x_i = 0$ to $x_1 = 1$, since (recall (9))

$$B_0 - c - \alpha_i c + \gamma_i \left(NB_0 - c\right) < 0. \tag{27}$$

At $t = 1$ 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 transformation of the interaction into a social dilemma is publicly observable. Hence, the personal norm switches from $y_i = 0$ to $y_i = 1$ for all $i \in I$ (recall (7)). Given the assumptions on how first-order 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 $\tilde{\gamma}(\gamma_i, B)$ (recall (10)). Hence,

$$n_1 = \#\mathcal{C}(I, \Theta), \tag{28}$$

which leads any individual i to hold the belief for $t = 2$ equal to

$$\hat{s}_{i,2} = \frac{n_1}{N-1}. \tag{29}$$

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

$$B - c + \gamma_i(NB - c) - \alpha_i c \left(1 - \frac{n_1}{N-1}\right) \geq -\beta_i \frac{c\, n_1}{N-1}, \tag{30}$$

or, equivalently,

$$\gamma \geq \frac{(1+\alpha_i)c - B - (\alpha_i + \beta_i)cn_1/(N-1)}{NB - c} = \bar{\gamma}\left(\alpha_i, \beta_i, \frac{n_1}{N-1}, B\right), \tag{31}$$

where the threshold $\bar{\gamma}(\alpha_i, \beta_i, s, B)$ was already defined in (20) for the purpose of the 2-type examples. Note now that $\tilde{\gamma}(\alpha_i, B)$, that is, the threshold value for the Kantian concern that defines the set of individuals who are the first to switch from $x_i = 0$ to $x_i = 1$, equals $\bar{\gamma}(\alpha_i, \beta_i, n_0/(N-1), B)$, where $n_0 = 0$, and that $\bar{\gamma}(\alpha_i, \beta_i, s, B)$ is decreasing in s. It follows that the total number of contributors at time $t = 2$ is

$$n_2 = \#\{i \in I \mid \gamma_i \geq \bar{\gamma}(\alpha_i, \beta_i, n_1/(N-1), B)\}. \tag{32}$$

More generally, at any point in time, any individual i holds the belief for period t equal to

$$\hat{s}_{i,t} = \frac{n_{t-1}}{N-1}, \tag{33}$$

and contributes at time t if and only if

$$\gamma_i \geq \bar{\gamma}(\alpha_i, \beta_i, n_{t-1}/(N-1), B). \tag{34}$$

The following equation is sufficient to describe the total number of contributors at any time $t \geq 0$:

$$n_t = \begin{cases} 0 \text{ if } t = 0 \\ \#\{i \in I \mid \gamma_i \geq \bar{\gamma}(\alpha_i, \beta_i, n_{t-1}/(N-1), B)\} \text{ if } t \geq 1. \end{cases} \tag{35}$$

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, the process stops (in the sense that $n_t = n_{\hat{t}}$ for all $t > \hat{t}$) within finite time. Figure 4 shows a possible dynamic. In the proposition below, we also show that the dynamic must stop when the

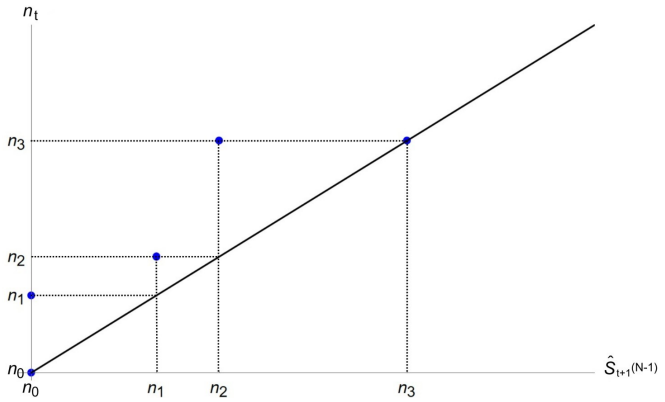

Figure 4: A possible dynamic with $n_0 = 0$ contributors at $t = 0$, n_1 at $t = 1$, etc. The straight line is the 45-degree line.

number of contributions reaches the smallest number of contributions associated with a Nash equilibrium of the static game $\Gamma = \langle (N, B, c), \Theta \rangle$.

Proposition 6. *Suppose that at time $t = 1$ the marginal benefit from contributing increases sufficiently to transform the interaction from a social non-dilemma to a social dilemma. Suppose that prior to the change no individual contributed, and that individuals are myopic. Then:*

1. *the increase in B generates some behavioral change at time $t = 1$ if and only if in the social dilemma there are some committed contributors, i.e., $\mathcal{C}(I, \Theta) \neq \emptyset$.*

2. *there exists a finite $\hat{t} \geq 1$ such that no further behavioral changes occur after time period $\hat{t}$;*

3. *for any $t \geq \hat{t}$, the number of contributors is the smallest one associated with a Nash equilibrium of the static game $\Gamma = \langle (N, B, c), \Theta \rangle$.*

Proof. The first two points were proven in the text. To prove the last point, let $\underline{n}$ denote the smallest number of contributors associated with a Nash equilibrium of the static game $\Gamma = \langle (N, B, c), \Theta \rangle$. If $\underline{n} = 0$, then $\mathcal{C}(I, \Theta) = \emptyset$ (by Proposition 4), implying $n_t = 0$ for all $t \geq 0$. Turning to the case where $\mathcal{C}(I, \Theta) \neq \emptyset$, so that $\underline{n} \geq 1$, define t' as the time

period at which the dynamic described by (35) reaches for the first time some number $n_{t'} \geq \underline{n}$ of contributors. I first show that this time period exists. To see this, suppose by contradiction that the dynamic stops at some time $t'' < t'$. That the dynamic stops means that $n_{t''+1} = n_{t''}$, which in turn implies

$$m((n_{t''} - 1)/(N - 1)) = n_{t''}, \tag{36}$$

where the function m is defined in (17). In words, there are exactly $n_{t''}$ individuals who prefer to contribute given that $n_{t''} - 1$ others do so. But this means that there exists a Nash equilibrium of the static game $\Gamma = \langle (N, B, c), \Theta \rangle$ at which $n^* = n_{t''}$. However, by definition of t', $n_{t''} < \underline{n}$, so that a contradiction with the definition of $\underline{n}$ is reached.

The second step of the proof consists in showing that $n_{t'} = \underline{n}$. By definition of t', $n_{t'-1} \leq \underline{n} - 1$, which in turn implies $\bar{\gamma}(\alpha_i, \beta_i, n_{t'-1}/(N-1), B) \geq \bar{\gamma}(\alpha_i, \beta_i, (\underline{n} - 1)/(N - 1), B)$, and hence $n_{t'} \leq \underline{n}$. Since, by definition of t', we have $n_{t'} \geq \underline{n}$, this implies $n_{t'} = \underline{n}$.

As a final step, recall that by definition of $\underline{n}$, there are exactly $\underline{n}$ individuals who prefer to contribute given that $\underline{n} - 1$ others do so. This, together with $n_{t'} = \underline{n}$, implies

$$n_{t'+1} = \#\{i \in I \mid \gamma_i \geq \bar{\gamma}(\alpha_i, \beta_i, n_{t'}/(N-1), B)\} = n_{t'}. \tag{37}$$

In other words, the process stops at t'. □

Figure 5 shows what these contribution dynamics would have been if $N = 95$ and the preference distribution had been the one in Table 1, for a cost $c = 1$ and four different values of the marginal benefit B; starting from the bottom line and moving upwards, $B = 0.05$, $B = 0.075$, $B = 0.1$, and $B = 0.15$. All the dynamics are stabilized after between two and four periods. The dynamics converge to $n^* = 68$ if $B = 0.15$, $n^* = 60$ if $B = 0.1$, $n^* = 38$ if $B = 0.075$, and $n^* = 14$ if $B = 0.05$ (which coincide with the unique Nash equilibrium of the static game in each case, recall Section 3). 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, in this illustrating example a higher B induces a higher share of the potential contributors to contribute.

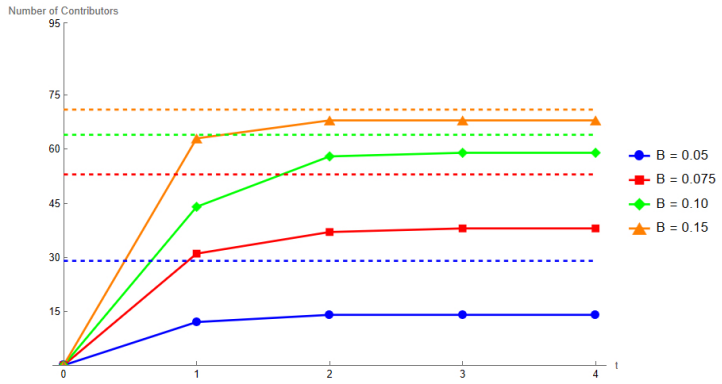

Figure 5: Contribution dynamics for $B = 0.05$ (bottom), $B = 0.075$ (second from bottom), $B = 0.1$ (third from bottom), and $B = 0.15$ (top); and $c = 1$. Left panel: the straight line is the 45-degree line. The dashed lines show the total number of potential contributors (committed and conditional contributors).

5.1.2 Heterogenous beliefs about the marginal benefit B

The assumption that individuals have correct beliefs about the marginal benefit is now discarded. 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, and hold the personal norm $y_i = 0$. Suppose further that initially no individual contributes. At some point in time, say $t = 1$, 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 $y_i = 1$. Hence, j contributes at time $t = 1$ if and only if $\gamma_j \geq \bar{\gamma}(\alpha_j, \beta_j, 0, B)$ (recall (31)). I will call *leaders* those who at time $t = 1$ are (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, the set of leaders is

$$\mathcal{L}(J) = \{j \in J \,|\, \gamma_j \geq \bar{\gamma}(\alpha_j, \beta_j, 0, B)\}, \tag{38}$$

so that

$$n_1 = \#\mathcal{L}(J). \tag{39}$$

Among the uninformed individuals the 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 exhibit a sufficiently strong solidarity towards the contributors. The analysis above shows that any uninformed individual k (who holds the belief $B_k = \hat{B}$) with empirical expectation $s_{k,t}$ contributes if and only if

$$\beta_k \geq \bar{\beta}(\alpha_k, \gamma_k, s_{k,t}, \hat{B}). \tag{40}$$

More informed individuals may also start contributing, having seen n_1 others doing so in $t = 1$. Hence, the total number of individuals who contribute at $t = 2$ is

$$n_2 = \#\{j \in J \mid \gamma_j \geq \bar{\gamma}(\alpha_j, \beta_j, n_1/(N-1), B)\} + \#\{k \notin J \mid \beta_k \geq \bar{\beta}(\alpha_k, \gamma_k, n_1/(N-1), \hat{B})\}. \tag{41}$$

This may in turn trigger further contributions among both the informed and the uninformed individuals. More generally, at any time $t \geq 1$, the number of contributors is:

$$n_t = \#\{j \in J \mid \gamma_j \geq \bar{\gamma}(\alpha_j, \beta_j, n_{t-1}/(N-1), B)\} + \#\{k \notin J \mid \beta_k \geq \bar{\beta}(\alpha_k, \gamma_k, n_{t-1}/(N-1), \hat{B})\}. \tag{42}$$

Ceteris paribus, $\bar{\gamma}$ and $\bar{\beta}$ are 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. Moreover, at any point in time there may be both informed and uninformed individuals who contribute. This proposition follows from arguments already developed above:

Proposition 7. *Suppose that the situation is a social dilemma, $NB > c > B$, but that at times $t \leq 0$ all individuals incorrectly believe that it is a social non-dilemma and thus hold the personal norm $y_i = 0$. At $t = 1$ 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, i.e., $\mathcal{L}(J) \neq \emptyset$.

3. *If at* $\hat{t}$ *there are still some individuals who do not contribute, these may be uninformed and/or informed.*

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

5.1.3 Correcting first-order beliefs

The assumption that past behaviors are public information, admittedly a strong assumption in most cases, is now dropped. Individuals are still assumed to be myopic, however, in the sense that in each period they best-respond to their first-order beliefs about behavior in the preceding period. Attention is restricted to social dilemmas (it would be straightforward to adapt the reasoning to social non-dilemmas) and the goal is to make three simple points.

Thus, consider a population where individuals have correct beliefs about the marginal benefit B, so that i's threshold is given by (recall 12):

$$\tilde{s}(\theta_i, B) = \begin{cases} 0 \text{ if } \gamma_i \geq \frac{(1+\alpha_i)c-B}{NB-c} \\ \frac{(1+\alpha_i)c-B-\gamma_i(NB-c)}{(\alpha_i+\beta_i)c} \text{ otherwise.} \end{cases} \tag{43}$$

Let $\hat{\mathbf{s}}_{\mathbf{0}} = (\hat{s}_{1,0}, \hat{s}_{2,0}, ..., \hat{s}_{N,0})$ denote the vector of initial first-order beliefs, at time $t = 0$, and that in period $t = 1$ all the individuals simply best-respond to these beliefs.

Given that individuals have correct beliefs about B, both the committed contributors and the committed non-contributors have a dominant strategy (recall Proposition 2): whether their beliefs are correct or not, the former contribute while the latter don't. The first simple point is that while their behavior is independent of their first-order beliefs, their subjective utilities are not. For example, with an initial underestimation of the number of contributions, a committed contributor who dislikes being behind others ($\alpha_i > 0$) will experience a rise in utility while a committed non-contributor who dislikes being ahead of others ($\beta_i > 0$) will experience a decline in utility, following a correction

of their first-order beliefs.

Turning to the conditional contributors, suppose first that their initial beliefs are pessimistic: $\hat{s}_{i,0} = 0$ for all i. At $t = 1$ they do not contribute. With an intervention correcting their beliefs at the end of $t = 1$, they would hold beliefs $\hat{s}_{i,2} = \#\mathcal{C}(I, \Theta)$, and the number of contributors at $t = 2$ would be (recall (32))

$$n_2 = \#\{i \in I \mid \gamma_i \geq \bar{\gamma}(\alpha_i, \beta_i, \#\mathcal{C}(I, \Theta)/(N-1), B)\}. \tag{44}$$

Now, if the initial beliefs were erroneous because individuals do not seek out this information, then the beliefs will remain at $\hat{s}_{i,t} = \#\mathcal{C}(I, \Theta)$ and the number of contributors will remain at $n_t = n_2$ forever. If so, and this is the second simple point made here, a new correction of beliefs will be needed to generate any further changes. Recalling now that with correct myopic beliefs the dynamic would stop at when it reaches the smallest number of contributors corresponding to a Nash equilibrium. Hence, and following the notation in Proposition 6, it would be necessary to correct the beliefs for $\hat{t}$ periods for this point to be reached.

For the third point, suppose that initially some but not all conditional contributors contribute, so that the number of contributors is some $N - \mathcal{N}(I, \Theta) > M > \mathcal{C}(I, \Theta)$. Now assume that individuals fall prey to false consensus, so that their first-order beliefs are correlated with their own behavior. Formally, each conditional contributor who contributes holds some belief $\hat{s}_{i,0} > (M-1)/(N-1)$, while each conditional contributor who does not contribute holds some belief $\hat{s}_{i,0} < (M-1)/(N-1)$. A belief-correcting intervention may then make non-contributors begin contributing, but it can also lead contributors to stop contributing. Depending on the strength of the bias, and the distribution of preferences, the intervention may either be successful, have a nil effect, or even backfire.

5.2 Social non-dilemmas

Consider a population where initially all individuals *falsely believe* that the interaction is a social dilemma: $\hat{B}_i = \hat{B}$ for all i is such that $N\hat{B} > c > \hat{B}$. Initially, they thus hold the personal moral norm $y_i = 1$. Suppose further that initially all individuals select the costly action: $n_0 = N$ (note that we thus assume that $\mathcal{N}(I, \Theta, \hat{\mathcal{B}}) = \emptyset$). Like in the

previous subsection, the beliefs of a set $J \subseteq I$ of individuals are corrected at $t = 1$, all of whom switch their personal norm to $y_i = 0$. Leaders, who at $t = 1$ are informed and also willing to alter their behavior given their belief that all the others select the costly action, must now have a sufficiently weak solidarity towards the others. Using $\mathcal{M}$ to denote the set of leaders in this situation:

$$\mathcal{M}(J) = \{j \in J \,|\, \beta_j < \bar{\beta}(\alpha_j, \gamma_j, 1, B)\}, \tag{45}$$

where $\bar{\beta}(\alpha_j, \gamma_j, s, B)$ is defined in (21). Hence, the number of individuals who select the costly action at $t = 1$ is

$$n_1 = N - \#\mathcal{M}(J). \tag{46}$$

Some uninformed individuals may also cease selecting the costly action if sufficiently many informed individuals have done so, and if they exhibit a sufficiently weak solidarity towards the contributors. The analysis above shows that an uninformed individual k (who holds the belief $B_k = \hat{B}$) with empirical expectation $s_{k,t}$ refrains from the costly action if and only if

$$\beta_k < \bar{\beta}(\alpha_k, \gamma_k, s_{k,t}, \hat{B}). \tag{47}$$

More informed individuals may also switch to the costless action, having seen n_1 others doing so in $t = 1$. Hence, the total number of individuals who contribute at $t = 2$ is:

$$\begin{aligned} n_2 \;=\; & N - \#\{j \in J \mid \beta_j < \bar{\beta}(\alpha_j, \gamma_j, n_1/(N-1), B)\} \\ & - \#\{k \notin J \mid \beta_k < \bar{\beta}(\alpha_k, \gamma_k, n_1/(N-1), \hat{B})\}. \end{aligned} \tag{48}$$

More generally, at any time $t \geq 1$, the number of contributors is:

$$\begin{aligned} n_t \;=\; & N - \#\{j \in J \mid \beta_j < \bar{\beta}(\alpha_j, \gamma_j, n_{t-1}/(N-1), B)\} \\ & - \#\{k \notin J \mid \beta_k < \bar{\beta}(\alpha_k, \gamma_k, n_{t-1}/(N-1), \hat{B})\}. \end{aligned} \tag{49}$$

Since $\bar{\gamma}$ is decreasing in n_{t-1}, any individual who switched from contributing to not contributing at some point in time, will never switch back to contributing. Moreover, at any point in time there may be both informed and uninformed individuals who contribute.

This proposition follows from arguments already developed above:

Proposition 8. *Suppose that the situation is a social non-dilemma, $c > NB > B$, but that at times $t \leq 0$ all individuals incorrectly believe that the costly action is the right thing to do and thus hold the personal norm $y_i = 1$. At $t = 1$ a set $J \subseteq I$ of individuals obtain the correct information about B. Suppose that prior to the change all individuals selected the costly action, 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 ($n_{\hat{t}} < N$) if, and only if, there is at least one leader, i.e., $\mathcal{M}(J) \neq \emptyset$.*

3. *If at $\hat{t}$ there are still some individuals who select the costly action, these may be uninformed and/or informed.*

The correct information must now reach at least one individual with a sufficiently weak solidarity towards the others for it to have any effect on behavior.

With the preference distribution in Table 1, for four different values of the marginal benefit B, the dynamics would almost immediately lead to zero contributions (only three individuals would contribute upon learning that $c > NB$, and these three individuals then also switch to not contributing).

5.3 Summing up: behavioral norms across time and space

I conclude this section by noting that the model can indeed explain norm variation across time and space, as announced in the introduction.

When it comes to temporal variation, the analysis in this section shows that spontaneous changes in the behavioral norm in a given population appear as long as individuals with sufficiently strong Kantian concerns experience a change in their beliefs at some point in time, and this leads sufficiently many individuals to change their behavior as well. This can help explain why norms concerning smoking indoors have changed drastically in some countries. While legislation against smoking indoors in restaurants and official buildings certainly contributed to reducing this behavior, it cannot fully explain

why the behavior also has become less common at private parties. Social-Kantian preferences, on the other hand, can help explain this: as scientific evidence on the effects of secondhand smoke mounted a few decades ago, the personal moral norms would have changed, and strongly Kantian smokers would have started to step outside to smoke. When this happened in freezing temperatures (and it did!), the material sacrifice would have been non-negligible. Individual with a sufficiently pronounced sense of solidarity towards those enduring such a sacrifice would then have followed suit, etc.

When it comes to spatial variation, the model shows that different preference distributions can yield different behavioral norms, even for the same distribution of beliefs about the marginal benefit B. In particular, in social dilemmas the number of individuals with strong enough Kantian concerns and with a strong enough sense of solidarity is critical for there to exist a behavioral norm that is congruent with the personal moral norms.

6 Discussion: field and lab experiments

In the past couple of decades a host of field experiments have allowed scholars to evaluate the effectiveness of various informational interventions related to norms. Areas of application in line with my model include energy and water conservation, recycling, public transportation usage, and tax compliance, which indeed are situations with social dilemmas. The model is relevant for two types of interventions: social comparisons, and correction of beliefs about the marginal benefit.[13] Here I discuss the findings of some such field experiments in light of the model. The aim is *not* to provide an overview, but rather to make a few observations to highlight how the model proposed here might be useful for the design and the interpretation of such experiments.

6.1 Social comparison interventions

Social comparison interventions provide individuals with feedback on own behavior and the behavior of others in some reference group. This amounts to a correction of first-order beliefs, discussed in Section 5.1.3. The model predicts that if prior to the intervention individuals underestimate (respectively overestimate) others' efforts, the intervention should

[13]Second-order beliefs about others' normative views being absent from the model, studies that document effects of correcting such beliefs are not discussed (e.g., Bursztyn et al. (2020)).

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; while such a "boomerang effect" has been observed in some field experiments on electricity consumption (e.g., Schultz et al. (2007), Allcott (2011)), other studies have found overall positive effects of social comparison interventions on electricity and water consumption (e.g., Allcott (2011), Ferraro & Price (2013), Schultz et al. (2016) Brandon et al. (2019)).

Such behavioral responses could of course be driven by conformity bias. However, why would the positive effects of social comparisons on electricity and water consumption be absent in other contexts, such as in the study on public transportation usage by Gravert & Olsson Collentine (2021)? One explanation could be that the cost (in terms of comfort and time) from switching from car to public transportation is much larger than the cost of slightly reducing in one's electricity or water consumption. My model then offers an additional explanation compared to theories based on pure conformity bias, since this cost also determines the sacrifice that individuals perceive if they switch while others don't; if individuals are averse to making greater sacrifice than others, the effect of the cost is thus amplified.

In sum, the model suggests that gathering information about attitudes towards differences in material payoffs and the perceived sacrifices should help obtain a better understanding of the mechanisms that drive behavioral changes (or non-changes). Such data could also be useful in the experimental design stage, since they may improve the accuracy of the hypothesized predictions.

6.2 The importance of personal moral norms

The model can help explain the observation of asymmetric responses to social comparisons in some studies: increased effort (on average) among those with pre-intervention below-average efforts combined with either small positive or nil effects on those with pre-intervention above-average efforts; see, e.g., Allcott (2011) for a study on electricity consumption, and Ferraro & Price (2013) for a study on water consumption. Such an asymmetric response could be explained by pure conformity bias only if first-order beliefs about others' behaviors are asymmetric. By contrast, in my model such asymmetries

arise if the preference distribution implies a larger number of committed contributors than committed non-contributors. This explanation appears to be 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.

According to my model, both the personal norm and the utility cost of deviating from this norm is determined by a universalization argument. Hence, it may be useful to collect information about the participants' habits of resorting to such universalization arguments.

6.3 Beliefs about the marginal benefit

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 my 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: (1) it corrects the belief about the own net material benefit ($\hat{B}_i - c$) upwards; (2) the individual's threshold is reduced; and (3) it can lead to an upward adjustment of the personal norm (recall equation (7)). By contrast, in models relying solely on conformity bias, only the first effect would be at work.

6.4 Measuring preferences

Collecting data about the components of the preferences in the model proposed here may help improve the design of information-based policy interventions, as well as the empirical analysis and interpretation of the results. In particular, to what extent can conditionally cooperative behavior in social dilemmas, such as climate change mitigating behaviors, be linked to attitudes towards making larger sacrifices than others (α in the model)? And to what extent can the sustainability of socially inefficient costly behaviors, such as female genital mutilation, be attributed to a strong sense of solidarity among individuals (β in the model)? And would knowledge about whether individuals apply universalization reasoning help predict their personal moral norms (κ in the model)? Methods similar to

the ones applied in some recent studies based on lab and survey experiments to estimate these attitudes could prove useful (Bruhin et al. (2018), van Leeuwen & Alger (2024)). However, survey-based methods will be needed to scale up data collection (Andre et al. (2024)).

6.5 (Un)-conditional cooperation in lab experiments

A large number of lab experiments have examined the dynamics of contributions in repeated public goods games, using designs similar to those in Fischbacher et al. (2001). One recurring pattern is clearly compatible with semi-Kantian preferences: typically, and using my terminology, some participants are conditional contributors, some are non-contributors, while some are unconditional contributors.

Another recurring pattern is that the number of contributions declines over time. This would be compatible with semi-Kantian preferences combined with over-optimistic first-order beliefs. Alternatively, initial first-order beliefs could be correct, and the decline in contributions over time could be driven by the additional presence of "far-sighted free-riders", who contribute only in the early rounds in order to trigger high contributions levels by the conditional cooperators (Engel & Rockenbach (2024)).

Social-Kantian preferences could further offer an explanation for why some individuals are willing to lead by example in sequential public goods games: they would be subjects with strong enough Kantian moral concerns (Eichenseer (2023)); note, however, that their behavior is also compatible with far-sighted free-riding.

In sum, the results reported above suggest that a better understanding of behavioral patterns in repeated and sequential public goods games might be obtained by considering social-Kantian preferences. If so, this would be in line with existing experimental work which has shown in other settings that the explanatory power is enhanced thanks to these preferences (Miettinen et al. (2020), van Leeuwen & Alger (2024)).

7 Concluding remarks

As evidence against the idea that the purely materially self-interested *Homo oeconomicus* is found in every human being accumulates, interest in non-monetary policy instruments is on the rise (see, e.g, Thaler & Sunstein (2008), Johansson-Stenman & Konow (2010),

Croson & Treich (2014), Bowles (2016), Nyborg et al. (2016)). In particular, it is increasingly recognized that humans are complex social animals, whose behavior is influenced by norms, driven by several factors. We propose a model with 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, which if sufficiently pronounced makes them condition their behavior on that of others. For any given distribution of preferences and beliefs, this model produces endogenously a unique distribution of individual thresholds for collectively desirable behavior in social dilemmas, and for collectively undesirable behavior in social non-dilemmas.

Needless to say, to capture the full complexity of norms and norm change, a number of other factors should be considered. In particular, field experiments have documented effects of second-order beliefs about what others deem appropriate (e.g., Schultz et al. (2007), Bursztyn et al. (2020)), image concerns (e.g., Gerber et al. (2008)), punishment (e.g., Brouwer et al. (2023)) group identity (e.g., Ehret et al. (2022)); 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. It will be important to understand whether and how these factors interact with the ones analyzed in this paper.

Ultimately, a fine understanding of how humans and human societies function in practice is necessary to design desirable and effective policies (Duflo (2017)).[14] It is hoped that the model in this paper will prove useful for the design of field experiments and policy interventions.

References

Acemoglu, D., & Jackson, M. O. (2015). History, Expectations, and Leadership in the Evolution of Social Norms. *Review of Economic Studies*, *82*(2), 423–456.

[14]Relatedly, several recent contributions highlight the potential importance of inter-disciplinary approaches; see Bicchieri, Dimant, et al. (2023), Alger et al. (2024), Andrighetto et al. (2024), Gelfand et al. (2024), and Heyes (2024).

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 moralispreference 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.

Andre, P., Boneva, T., Chopra, F., & Falk, A. (2024). Globally representative evidence on the actual and perceived support for climate action. *Nature Climate Change*, *14*(3), 253–259.

Andreoni, J., Nikiforakis, N., & Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences*, *118*(16), e2014893118.

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). University of Rome Tor Vergata.

Bénabou, R., & Tirole, J. (in press). Laws and norms. *Journal of Political Economy*.

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

Besley, T., Jensen, A., & Persson, T. (2023). Norms, enforcement, and tax evasion. *Review of Economics and Statistics*, *105*(4), 998–1007.

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., Gächter, S., & Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, *132*, 59–72.

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

Bicchieri, C., & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, *22*(2), 191–208.

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). Linear Social Interactions Models. *Journal of Political Economy*, *123*(2), 444–496.

Blume, L., & Durlauf, S. (2003). 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). Youve 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.

Brandon, A., List, J. A., Metcalfe, R. D., Price, M. K., & Rundhammer, F. (2019). Testing for crowd out in social nudges: Evidence from a natural field experiment in the market for electricity. *Proceedings of the National Academy of Sciences*, *116*(12), 5293–5298.

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.

Brouwer, T., Galeotti, F., & Villeval, M. C. (2023). Teaching Norms: Direct Evidence of Parental Transmission. *Economic Journal*, *133*(650), 872–887.

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.

Capraro, V., & Rand, D. G. (2018). Do the right thing: Preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making*, *13*(1), 99–111.

Cardenas, J. C. (2011). Social Norms and Behavior in the Local Commons as Seen

Through the Lens of Field Experiments. *Environmental and Resource Economics*, *48*(3), 451–485.

Carlsson, F., Johansson-Stenman, O., & Nam, P. K. (2015). Funding a new bridge in rural Vietnam: a field experiment on social influence and default contributions. *Oxford Economic Papers*, *67*(4), 987–1014.

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. *Annual Review of Psychology*, *55*, 591–621.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology*, *24*, 201–234.

Congdon Fors, H., Isaksson, A.-S., & Lindskog, A. (2024). Changing local customs: The long run impacts of Christian missions on female genital cutting in Africa. *Journal of Development Economics*, *166*, 103180.

Croson, R., & Treich, N. (2014). Behavioral Environmental Economics: Promises and Challenges. *Environmental and Resource Economics*, *58*(3), 335–351.

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.

Dimant, E., Gelfand, M., Hochleitner, A., & Sonderegger, S. (2024). Strategic behavior with tight, loose, and polarized norms. *Management Science*.

Duflo, E. (2017). The economist as plumber. *American Economic Review*, *107*(5), 1–26.

Efferson, C., Ehret, S., Von Flüe, L., & Vogt, S. (2024). When norm change hurts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1897).

Efferson, C., Vogt, S., & Fehr, E. (2020). The promise and the peril of using social influence to reverse harmful traditions. *Nature Human Behaviour*, *4*(1), 55–68.

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.

Eichenseer, M. (2023). Leading-by-example in public goods experiments: What do we know? *Leadership Quarterly*, *34*(5), 101695.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, *3*(4), 99117.

Engel, C., & Rockenbach, B. (2024). What makes cooperation precarious? *Journal of Economic Psychology*, *101*, 102712.

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.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.

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*, 1–22.

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*(1897).

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.

Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, *102*(1), 33–48.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, *24*(3), 153–172.

Glaeser, E., & Scheinkman, J. A. (2000). *Non-market interactions* (WP8053). NBER.

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

Gravel, N., Laslier, J.-F., & Trannoy, A. (2000). Consistency between tastes and values: a universalisation approach. *Social Choice and Welfare*, *17*, 129—137.

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.

Hedström, P. (2005). *Dissecting the Social: On the Principles of Analytical Sociology.* Cambridge: Cambridge University Press.

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

Johansson-Stenman, O., & Konow, J. (2010). Fair air: Distributive justice and environmental economics. *Environmental and Resource Economics*, *46*(2), 147–166.

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

Krupka, E. L., & Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, *30*(3), 307–320.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524.

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

La Ferrara, E. (2019). Presidential address: Aspirations, social norms, and development. *Journal of the European Economic Association*, *17*(6), 1687–1722.

Laffont, J.-J. (1975). Macroeconomic constraints, economic efficiency and ethics: An introduction to kantian economics. *Economica*, *42*(168), 430–437.

Lane, T., Nosenzo, D., & Sonderegger, S. (2023). Law and norms: Empirical evidence. *American Economic Review*, *113*(5), 1255–1293.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(42), 26158–26169.

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 & Organization*, *173*, 1-25.

Mittal, D., López, F. G.-N., Constantino, S., Shalvi, S., Chen, X., & Vasconcelos, V. V. (2025). *Targeting heuristics for cost-optimized institutional incentives in heterogeneous networked populations.* Retrieved from `https://arxiv.org/abs/2501.13623`

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. *Annual Review of Resource Economics*, *10*, 405-423.

Nyborg, K., Anderies, J. M., Dannenberg, A., Lindahl, T., Schill, C., Schlüter, M., ... de Zeeuw, A. (2016). Social norms as solutions. *Science*, *354*(6308), 42-43.

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

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, *14*(3), 137–158.

Roemer, J. E. (2019). *How We Cooperate*. New Haven: Yale University Press.

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

Schram, A., Zheng, J. D., & Zhuravleva, T. (2022). Corruption: A cross-country comparison of contagion and conformism. *Journal of Economic Behavior & Organization*, *193*, 497-518.

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.

Szekely, A., Lipari, F., Antonioni, A., Paolucci, M., Sánchez, A., Tummolini, L., & Andrighetto, G. (2021). Evidence from a long-term experiment that collective risks change social norms and promote cooperation. *Nature Communications*, *12*(1), 1–7.

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.

Thaller, A., Fleiß, E., & Brudermann, T. (2020). No glory without sacrifice drivers of climate (in)action in the general population. *Environmental Science & Policy*, *114*, 7–13.

Thöni, C. (2014). Inequality aversion and antisocial punishment. *Theory and Decision*, *76*(4), 529–545.

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.

Vriens, E., Andrighetto, G., & Tummolini, L. (2024). Risk, sanctions and norm change: the formation and decay of social distancing norms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1897).

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			

The values in the table are obtained with the estimates of the behindness aversion, aheadness aversion, and Kantian concern parameters in van Leeuwen & Alger (2024) (called α_i, β_i, and κ_i in their article) when they posit risk-neutral subjects without reciprocity (i.e., the parameters δ_i and γ_i in their equation (1) are set to 0). These estimates are then divided by $(1-\kappa_i)$ (recall Footnote 10). To be in line with the assumptions, attention is restricted to the 95 subjects among the 112 core subjects in van Leeuwen & Alger (2024) for whom $\alpha_i + \beta_i \geq 0$.