Supplementary Information for "Tipping pro-environmental norm diffusion at scale: Opportunities and limitations"

Joël Berger[1,2,*], Charles Efferson[3,*], and Sonja Vogt[3,4,*]

[1]Department of Business and Management,
Kalaidos University of Applied Sciences, Switzerland

[2]Institute of Sociology, University of Bern, Switzerland

[3]Faculty of Business and Economics, University of Lausanne, Switzerland

[4]Nuffield College, University of Oxford, UK

[*]Address correspondence to joel.berger@soz.unibe.ch,
charles.efferson@unil.ch, and sonja.vogt@unil.ch.

1 A threshold distribution from Fehr-Schmidt preferences

Consider the game in Table S1.

Table S1. A two-person two-action public goods game with heterogeneous inequity aversion based on Fehr and Schmidt (1999). Individual i is the row player, and i' is the column player. See Fig. A1 from main paper for additional details.

	C	D
C	$(b-c, b-c)$	$\left(\frac{b}{2}-c-\alpha_i c, \frac{b}{2}-\beta_{i'} c\right)$
D	$\left(\frac{b}{2}-\beta_i c, \frac{b}{2}-c-\alpha_{i'} c\right)$	$(0,0)$

If i believes her partner will choose C with probability $\hat{q}_i$, and Π_i is a random variable for

i's payoff as a function of choice,

$$\begin{aligned} E[\Pi_i(\mathrm{C})] &= \hat{q}_i(b-c) + (1-\hat{q}_i)\left(\frac{b}{2} - c - \alpha_i c\right) \\ E[\Pi_i(\mathrm{D})] &= \hat{q}_i\left(\frac{b}{2} - \beta_i c\right) + (1-\hat{q}_i)(0). \end{aligned} \tag{1}$$

A threshold, q_i^*, is a belief that ensures i is indifferent between C and D, i.e. $\{q_i^* \in [0,1] \mid q_i^* = \hat{q}_i \text{ and } E[\Pi_i(\mathrm{C})] = E[\Pi_i(\mathrm{D})]\}$. $E[\Pi_i(\mathrm{C})] = E[\Pi_i(\mathrm{D})]$ implies

$$q_i^* = \frac{2c(1+\alpha_i) - b}{2c(\alpha_i + \beta_i)}. \tag{2}$$

In addition, $q_i^* \in [0,1]$ implies that $\alpha \geq (b-2c)/(2c)$ and $\beta \geq (2c-b)/(2c)$. The calculations for i' are identical. More generally, a distribution of (α, β) values subject to the constraints specified here produces a distribution of thresholds.

2 Threshold model with distorted beliefs and a beliefs-based intervention

We use q^* for thresholds and $\hat{q}$ for beliefs. A given individual cooperates if her belief about the probability a partner will cooperate is greater than or equal to her threshold. Let $f : [0,1]^2 \to \mathbb{R}$ be a density function specifying the distribution of thresholds and beliefs in the population before intervention. F is the joint cumulative distribution function. Let q_t be the proportion of the population cooperating at time t. Assume that before the intervention, i.e. $t = 0$, the population is in equilibrium in the sense that neither beliefs nor behaviors are evolving. Thus, $q_0 = \int_0^1 \int_0^{\hat{q}} f(q^*, \hat{q})\, \mathrm{d}\, q^*\, \mathrm{d}\, \hat{q}$ is a stable proportion of cooperation before intervention. A beliefs-based intervention means the social planner introduces some mechanism that ensures everyone always knows q_t. The intervention, in effect, supplants the potentially distorted process that governs how people formed beliefs about others before intervention. To see what effects such a mechanisms can have, we partition the domain of f, i.e. the unit square, as in Fig. S1.

$\underline{A}$ includes both its lower and upper boundaries. We denote the total mass as $\underline{x}_1$, where

$$\underline{x}_1 = \int_0^{q_0} \int_0^{\hat{q}} f(q^*, \hat{q})\, \mathrm{d}\, q^*\, \mathrm{d}\, \hat{q}. \tag{3}$$

$\overline{A}$ excludes its lower boundary, and it includes its right boundary. The total mass is $\overline{x}_1$, where

$$\overline{x}_1 = \int_0^1 \int_0^{q_0} f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q} - \int_0^{q_0} \int_0^{q_0} f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q}. \tag{4}$$

To draw a link with the main paper, $A = \underline{A} \cup \overline{A}$, and $x_1 = \underline{x}_1 + \overline{x}_1$.

B excludes its left boundary, but it includes its lower boundary. The total mass is x_2, where

$$x_2 = \int_{q_0}^1 \int_0^{\hat{q}} f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q} - \int_{q_0}^1 \int_0^{q_0} f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q}. \tag{5}$$

C excludes its upper boundary and includes its right boundary. The total mass is x_3, where

$$x_3 = \int_0^{q_0} \int_0^{q_0} f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q} - \int_0^{q_0} \int_0^{\hat{q}} f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q}. \tag{6}$$

$\underline{D}$ excludes both its left and upper boundaries. The total mass is $\underline{x}_4$, where

$$\underline{x}_4 = 1 - \int_{q_0}^1 \int_0^1 f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q} - \left(\int_0^1 \int_0^{q_0} f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q} - \int_{q_0}^1 \int_0^{q_0} f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q} \right). \tag{7}$$

$\overline{D}$ excludes its upper boundary but includes its lower boundary. The total mass is $\overline{x}_4$, where

$$\overline{x}_4 = \int_{q_0}^1 \int_{q_0}^1 f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q} - \int_{q_0}^1 \int_{q_0}^{\hat{q}} f(q^*, \hat{q}) \, \mathrm{d}\, q^* \, \mathrm{d}\, \hat{q}. \tag{8}$$

To draw a link with the main paper, $D = \underline{D} \cup \overline{D}$, and $x_4 = \underline{x}_4 + \overline{x}_4$.

To understand the value of the partition, recall one of the key assumptions of the threshold model (Granovetter, 1978), an assumption that we retain in modified form. Namely, the threshold model assumes that each individual chooses the behavior in question, in our case cooperate, in $t + 1$ if the proportion choosing the behavior in t was greater than or equal to the individual's threshold. The weak inequality used here is precisely what allows one to use the cumulative distribution function as the map for cultural evolutionary dynamics. We retain the assumption in modified form by assuming that each individual chooses to cooperate if her beliefs about others cooperating is greater than or equal to her threshold. Beliefs can be inaccurate before intervention, for whatever reason. The intervention makes the distribution of behaviors in t public knowledge, and each individual takes q_t as her belief about cooperation in $t + 1$. Thus, belief formation after intervention is based on accurate information, but it is not forward looking. Instead, it is myopic, and people best respond given

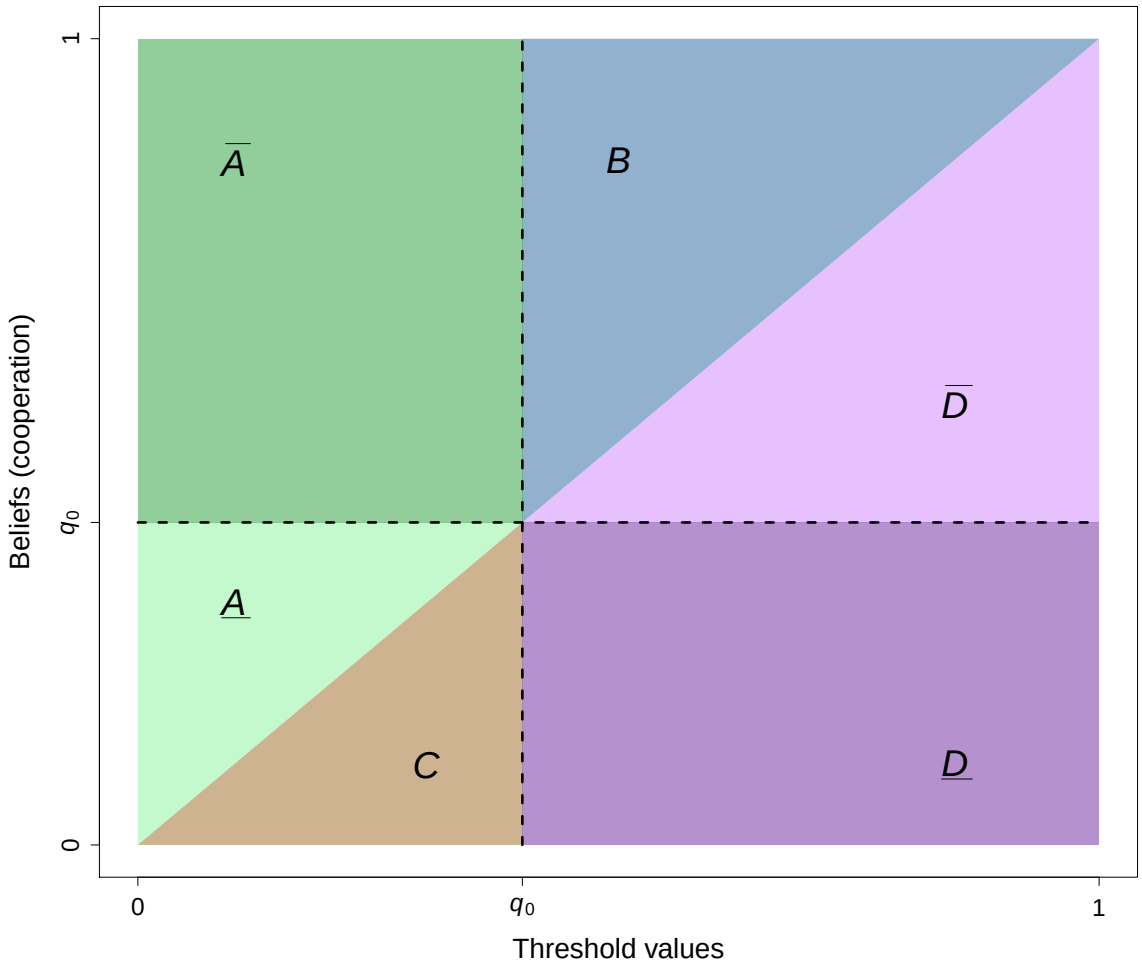

Figure S1. A partition of the threshold-belief space. As explained in the supplementary text, the mass of individuals in region B and the mass of individuals in region C exclusively determine the immediate effect of a beliefs-based intervention. Subsequent to this immediate effect, the cultural evolution of cooperation proceeds according to marginal threshold distribution, which accounts for the distribution of individuals over all six regions. The immediate effect reveals the direction but not the magnitude of the long-run effect.

these beliefs, as in the original threshold model (Granovetter, 1978). This kind of myopic best response is consistent with an extraordinary proportion of choices in experimental studies with coordination games (Mäs and Nax, 2016; Andreoni *et al.*, 2021). In particular, it means that we treat the beliefs-based intervention as a type of intervention that recovers the original threshold model, a point to which we return later.

Before doing so, however, let us examine the initial effect of the intervention by considering the effect on each of the subsets in our partition. Because $\underline{A}$ includes its lower boundary, the individuals in this subset cooperate before intervention. More generally, $\underline{A}$ includes individuals who, before intervention, underestimate the rate of cooperation in the population

because their beliefs are (weakly) too low. Despite the fact that they (weakly) underestimate the rate of cooperation, they cooperate because their beliefs are greater than or equal to their thresholds. The weak inequality here is equivalent to saying that $\underline{A}$ includes its lower boundary. These individuals would also cooperate if they had accurate beliefs because the intervention will either leave beliefs unchanged (i.e. the upper boundary of $\underline{A}$) or increase beliefs. Thus, the intervention will not lead individuals in $\underline{A}$ to change their behavior as an immediate consequence of the intervention.

Because $\overline{A}$ excludes its lower boundary, the individuals in $\overline{A}$ strictly overestimate the degree of cooperation in the population before intervention. Because their beliefs are greater than their thresholds, they cooperate before intervention. Moreover, because $\overline{A}$ includes its right boundary, they would also cooperate even if they had accurate beliefs. Thus, the individuals in $\overline{A}$ cooperate before intervention, and the intervention will not lead them to an immediate change in behavior.

B includes individuals who, before intervention, overestimate the degree of cooperation. Moreover, B includes its lower boundary, and so all the individuals in this subset cooperate because their beliefs are greater than or equal to their thresholds. However, because B excludes its left boundary, these individuals would not cooperate if they had accurate beliefs. Thus, the individuals in B cooperate specifically because of their optimistic beliefs, and the immediate effect of the intervention is to lead them to switch to defection.

C includes individuals who, before intervention, underestimate the degree of cooperation. In addition, because C excludes its upper boundary, the individuals in C all have beliefs strictly smaller than their respective thresholds, and so they do not cooperate. However, because C includes its right boundary, the individuals in the subset would cooperate if they had accurate beliefs. In this sense, C consists of individuals who defect specifically because of their pessimistic beliefs, and the immediate effect of the intervention is to lead them to switch to cooperation.

$\underline{D}$ includes individuals who, before intervention, underestimate the degree of cooperation, and they do not cooperate because their beliefs are strictly less than their thresholds. Moreover, because $\underline{D}$ excludes its left boundary, no one in $\underline{D}$ would cooperate if they had accurate beliefs. Thus, $\underline{D}$ consists of individuals who defect before intervention and do not change behavior as an immediate consequence of the intervention. $\overline{D}$ includes individuals who overestimate the degree of cooperation before intervention. Nonetheless, because $\overline{D}$ excludes its

upper boundary, these individuals all have thresholds strictly greater than their beliefs, and so they defect before intervention. No one in the subset would cooperate if they had accurate beliefs. Thus, $\overline{D}$ consists of individuals who defect before the intervention and do not change behavior as an immediate consequence of the intervention.

Notice that the only individuals to change their behavior as an immediate response to the beliefs-based intervention are individuals in B, who switch from cooperating to defecting, and individuals in C, who switch from defecting to cooperating. As a result, let q_1 be the proportion of individuals cooperating in $t = 1$. We designate $t = 1$ as the point in time when individuals who were cooperating or defecting specifically because of distorted beliefs have changed their behaviors as an immediate response to the intervention, but no additional cultural evolutionary dynamics have yet occurred. This means, in effect, that

$$q_1 = q_0 - x_2 + x_3. \tag{9}$$

The immediate effect of ensuring correct beliefs can thus be positive, negative, or neutral. If $x_2 \leq x_3$, $q_1 \geq q_0$, and cooperation (weakly) rises. If $x_2 > x_3$, $q_1 < q_0$, and cooperation declines. Which of these scenarios holds will depend on the distribution of individuals in regions B and C. The distribution of individuals in regions A and D is irrelevant in terms of the *immediate* effect of the intervention.

What happens after the immediate effect? To answer this question, note that the intervention represents a fundamental change in the informational setting. It eliminates the possibility of distorted beliefs by making the current rate of cooperation public knowledge at all points in time. As a result, beliefs become accurate, subject to the assumption of myopic updating, and they are always the same for everyone. Once this happens, the preference heterogeneity represented by the distribution of thresholds is the mechanism driving the cultural evolution of cooperation. This is true in terms of both the immediate effect of the intervention, namely the transition from $t = 0$ to $t = 1$, and it remains true for subsequent points in time. Technically, $\forall t \in \{1, 2, \ldots\}$, $q_t = G_(q_{t-1})$, where G is the marginal cumulative distribution of thresholds,

$$\begin{aligned} G(q^*) &= \int_0^{q^*} \int_0^1 f(q^*, \hat{q}) \, \mathrm{d}\, \hat{q} \, \mathrm{d}\, q^* \\ &= F(q^*, 1). \end{aligned} \tag{10}$$

A steady state of the resulting system (Granovetter, 1978; Efferson *et al.*, 2020) is a distribution of behavior in the population, $\tilde{q}$, such that $\tilde{q} = G(\tilde{q})$. The local stability of the steady state can be examined via cobwebbing (Granovetter, 1978) or via the standard techniques used for local stability analyses of nonlinear systems in discrete time (e.g. Hoy *et al.*, 2001). In particular, if $|G'(\tilde{q})| < 1$, $\tilde{q}$ is locally stable. Because G is a cumulative distribution function and thus monotone, this condition reduces simply to $G'(\tilde{q}) < 1$.

In general terms, a beliefs-based intervention takes a system defined in terms of heterogeneous preferences and heterogeneous - potentially distorted - beliefs and transforms it into a system defined only in terms of heterogeneous preferences. When this happens, under myopic best responses, the cultural evolution of cooperation proceeds according to the classic threshold model of Granovetter (1978), with G as the relevant distribution function. In this sense, when viewing the problem in terms of the *original* distribution of individuals, before intervention, all six regions of Fig. 1 are relevant for long-run cultural evolution. This is true simply because all six regions are necessary to give us the marginal distribution, G, that matters after intervention.

That said, the initial effect of the intervention depends entirely on the relative masses of regions B and C. Moreover, this initial effect also gives the direction of the long-run effect of the beliefs-based intervention, even if it does not tell us the magnitude of the long-run effect. To see this, consider the case in which $x_2 > x_3$. The intervention has an initially negative effect. Because G becomes relevant as soon as the intervention is introduced, $q_1 = q_0 - x_2 + x_3 = G(q_0) < q_0$. In addition, because G is a cumulative distribution function and thus monotonically increasing, and because we are focusing on the case in which $q_1 < q_0$, then $q_2 = G(q_1) \leq q_1 = G(q_0) < q_0$. By extension, for any $t > 1$, $q_t \leq q_1 < q_0$. Thus, if the immediate effect of the intervention is to provoke a decline in cooperation, then cooperation will not go back up. The rate of cooperation when the population stabilizes will depend on the shape of G and potentially the starting point, q_0. Whatever the final details, if the initial movement is downward, the long-run effect of the intervention will be negative.

Analogously, consider $x_2 < x_3$. In this case, $q_1 = q_0 - x_2 + x_3 = G(q_0) > q_0$. For any $t > 1$, $q_t \geq q_1 > q_0$. Again, the population may stabilize on a small increase in cooperation or a large increase. This will depend on the overall shape of G and potentially q_0. Regardless, if the initial movement is upward, the long-run effect of the intervention will be positive.

References

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

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

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

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

Hoy, M., Livernois, J., McKenna, C., Rees, R., and Stengos, T. (2001). *Mathematics for Economics*. Cambridge: The MIT Press, 2nd edition.

Mäs, M. and Nax, H. H. (2016). A behavioral study of "noise" in coordination games. *Journal of Economic Theory*, **162**, 195–208.