PNAS
www.pnas.org

Supplementary Information for

Predicting social tipping and norm change in controlled experiments

James Andreoni, Nikos Nikiforakis[*], Simon Siegenthaler

[*] Corresponding Author: Nikos Nikiforakis, New York University Abu Dhabi, P.O. Box 129188, Abu Dhabi, United Arab Emirates, Phone: +971 (0)26285436, Email: nikos.nikiforakis@nyu.edu

This PDF file includes:

Supplementary Text
Figures S1 to S10
Legend for Experimental Instructions S1
Legend for Dataset S1
Legend for Data Analysis S1
Legend for Model Simulation S1
SI References

Other supplementary materials for this manuscript include the following:

Experimental Instructions S1
Dataset S1
Data Analysis S1
Model Simulation S1

1. Materials and Methods

Subject Recruitment

The experiment was conducted at the economics laboratory of the University of California, San Diego (UCSD). The experimental protocol was approved by the IRB at NYU Abu Dhabi (#049-2016) and the IRB at UCSD (#150689). Informed consent was obtained and participants were informed that they could terminate their participation at any point in the experiment. We used the experimental software z-Tree (https://www.ztree.uzh.ch). Recruitment was done via the recruitment platform ORSEE (http://www.orsee.org). When signing up for the experiment, subjects only knew that they will participate in a decision-making experiment; the experiment was explained in detail upon subjects' arrival at the lab. A total of 54 sessions were run with 1,020 subjects. Each subject participated in one session only. All sessions consisted of 20 subjects, except for one experimental condition, which had 10 subjects per session. Subjects were students at UCSD from various disciplines. The mean age was 20 years and 54% of the participants were female.

Subject Experience During the Experiment

Upon arriving at the laboratory, written instructions on how to make decisions in the experiment were distributed to the subjects, which the experimenter also read aloud. The experiment started once all subjects had correctly answered a number of comprehension questions included at the end of the instructions.

Subjects interacted via computer terminals. At the start of each of 31 periods, subjects were told their "type". Types determined a subject's preferences over two alternative choices: Blue and Green. Specifically, subjects of type A received higher individual financial rewards for choosing Blue, while subjects of type B received higher individual financial rewards for choosing Green. At the start of the experiment, the reward for Blue exceeded the one for Green for all subjects. Over time, subjects' preferences changed gradually at a commonly known rate of 10% (i.e., subjects gradually switched from type A to type B). This change in preferences was explained in the instructions and, hence, was public knowledge. After learning their type in a given period, subjects were matched into pairs and were asked to choose between actions Blue and Green.

If two matched subjects chose different colors (i.e., they did not coordinate), their financial reward was reduced. The penalty depended on the number of people in the session choosing the other color. This created an incentive to conform to the majority choice. We refer to the Experimental Instructions (separate file) for a complete description of the experiment. However, for convenience, we reproduce below (in italics) the part of the instructions from the baseline treatment *TT-43* pertaining to subjects' incentives:

A Type A participant receives ***30 ECU*** [Experimental Currency Units] *when they choose BLUE and* ***20 ECU*** *when they choose GREEN. A Type B participant receives* ***20 ECU*** *when they choose BLUE and* ***30 ECU*** *when they choose GREEN. If the other participant chooses the same color, these are the earnings in a given round.*

However, every time you and the other participant choose ***different colors*** *you both receive a "****miscoordination penalty****". The penalty may differ for the two participants. In particular, the amount you will receive will be* ***reduced by 4 ECU for each participant in your matching group (i.e., the group of 20 participants) that chose a different color than you****. That is, the more people choose a different action than you, the greater will be your miscoordination penalty.*

At the end of each period, subjects received feedback. They could see their earnings and the number of subjects in the group choosing Blue and Green in the previous period. They were also informed about the choice of the specific subject they were matched with in the current period. Then, a new period began in which subjects were randomly re-matched. The central trade-off subjects faced was between their changing individual preferences from Blue to Green (their desire for change) and the cost of deviating from the "Blue norm" (the pressure to conform), which was initially established because everyone preferred Blue at the start of the experiment. The game ended after period 31.

After the main experiment, we continued by eliciting subjects' risk and nonconformity preferences. In the risk elicitation task, subjects had to pick one of six lotteries: (a) 8 in 10 chance to win $2, (b) 7 in 10 chance to win $3, (c) 6 in 10 chance to win $4, (d) 5 in 10 chance to win $5, (e) 4 in 10 chance to win $6, and (f) 3 in 10 chance to win $7. Options (a) to (f) order subjects by risk aversion, with (a) revealing the greatest risk aversion, (d) revealing risk neutrality (it maximizes expected value), and (f) is the most risk loving choice. The distribution of lottery choices is almost identical between the different treatments, indicating that there is no treatment effect on elicited risk preferences (see *Data Analysis S1*).

To elicit nonconformity preferences, subjects had to rate statements taken from a scale discussed in *(1)*. A five-point rating scale from 1 (strongly disagree) to 5 (strongly agree) was used. The statements were: *I become angry when my freedom of choice is restricted; It disappoints me to see others submitting to standards and rules; When someone forces me to do something, I feel like doing the opposite; I become frustrated when I am unable to make free and independent decisions; I find contradicting others stimulating. Regulations trigger a sense of resistance in me; The thought* *of being dependent on others aggravates me; It irritates me when someone points out things which are obvious to me; I am content only when I am acting of my own free will; I resist the attempts of others to influence me.* The distribution of scores in the nonconformity elicitation task is almost identical between the different treatments, indicating that there is no treatment effect on elicited conformity preferences (see *Data Analysis S1*).

At the end of a session, subjects were privately paid in cash. All rounds of the experiment were paid. The accumulated ECUs were exchanged to USD at a rate of 1 ECU = 0.03 USD. Everyone received \$10 as an initial budget. Subjects also received between \$0 and \$7 from the lottery task and \$3 for completing the survey on nonconformity preferences. If a subject made losses during the experiment, these were subtracted from the initial budget and the earnings in the lottery and nonconformity task. If a subject's earnings were below \$0 at the end of a session, the subject received \$0. Only 4 of the 1,020 subjects earned \$0. Payments averaged \$36.1 per subject. Sessions lasted less than 75 minutes.

Experimental Conditions

Our social-tipping model predicts that the likelihood of observing change depends on (*i*) the tipping threshold, which in turn depends on the benefit-cost ratio of norm change (v/p) and (*ii*) individual-specific preferences and expectations about the likelihood of change and their contribution to it (captured through γ_i). To provide a comprehensive test of these hypotheses, we implemented 9 experimental conditions. The corresponding instructions that were distributed to the subjects can be found in the separate file "Experimental Instructions". A description of the experimental conditions follows.

The first four conditions vary the tipping threshold to study how the benefit-cost ratio of norm abandonment affects the probability of social tipping, and whether societies can lower social penalties sufficiently if given the opportunity to do so.

Conditions varying the tipping threshold

TT-43 (Baseline): In each session, 20 subjects are randomly matched into pairs in each period and choose between two alternative actions: Blue or Green. Initially, everyone prefers Blue. In each period, each individual who has not previously switched to preferring Green has a 10% probability that his or her preference switches from Blue to Green. Choosing the preferred color yields a payoff of $v_H = 30$ ECUs and choosing the other color a payoff of $v_L = 20$ ECUs. Hence, the marginal benefit from change is $v = 30 - 20 = 10$ ECUs (corresponding to \$0.3) per period. In case subjects fail to choose the same color, they incur a miscoordination penalty of 4 ECU for each subject in the group choosing the other color. The maximum penalty is thus $p = 76$ ECUs (19 subjects choosing the other color times a penalty of 4 ECUs per subject). These parameters imply a tipping threshold of 43%, as $f_{TT} = 0.5 - 0.5\ v/p = 0.43$, which is above the theoretical cutoff for tipping. Hence, we predict no tipping and detrimental norm persistence for this condition. At the end of a period, subjects are informed about the action chosen by their matched subject but not about other subjects' preferences/types. With a delay of one period, subjects are also informed about their earnings and the total number of group members who chose Blue and Green. The game ends after period 31.

TT-30: Implements the same setting as in *TT-43* except that the benefit of choosing Green for a subject preferring Green is increased from 30 ECUs to $v_H = 50$ ECUs (the benefit of choosing Blue for a subject preferring Blue remains 30, and $v_L = 20$). Hence, the marginal benefit from change equals $v = 50 - 20 = 30$ ECUs (corresponding to 0.9$) per period. The tipping threshold in this condition is at 30%, i.e., below the theoretical cutoff for tipping of 35% (see *Fig. 2* in the article).

TT-23: Implements the same setting as in *TT-43* except that the miscoordination penalty per subject choosing the opposite color is reduced from 4 ECUs to 1 ECU. The maximum penalty is therefore $p = 19$ ECUs (19 subjects choosing the other color times a penalty of 1 per subject). The tipping threshold in this condition is at 23%, i.e., well below the theoretical cutoff for tipping (see *Fig. 2* in the article).

TT-Endo: Implements the same setting as in *TT-43* except that, in each period, subjects choose the miscoordination penalty their matched subject incurs per subject in the group choosing the other color. The available choices are a miscoordination penalty of 1 ECU as in *TT-23*, of 4 ECUs as in *TT-43*, or of 7 ECUs to not artificially bias penalties below those in the baseline condition. The color choice and the penalty choice are made simultaneously, before being informed about the behavior of the matched subject. In this condition, the penalties and therefore the tipping threshold are endogenous; the tipping threshold lies between 23% if everyone chooses a penalty of 1 and 46% if everyone chooses a penalty of 7. Similarly, the maximum penalty p is between 19 and 76.

Conditions varying social expectations and incentives for instigating change

The remaining five conditions keep the tipping threshold constant at the baseline level of 43% to study how expectations and incentives to lead change affect the probability of social tipping.

Fast Feedback: Implements the same setting as in *TT-43* except that subjects immediately learn at the end of each period how many others in the group chose Blue or Green. In particular, the one-period information delay present in *TT-43* is eliminated.

Small Society: The group size is halved compared to *TT-43*, from 20 to 10 subjects. To keep the tipping threshold identical to *TT-43*, the penalty per subject choosing the other color is increased from 4 to 8.44 ECUs. This keeps the maximum penalty at $p = 76$, identical to the baseline condition. Suppose one player chooses Green while everyone else chooses Blue. In *TT-43* the total penalty incurred by this player is 19 players times 4 ECUs, which equals 76 ECUs. In *Small Society* the total penalty is the same, 9 players times 8.44 ECUs, which also equals 76 ECUs. Thus, this treatment allows us to study how group size affects the probability of tipping, holding everything else constant.

Public Awareness: Implements the same setting as in *TT-43.* The only difference is that, in the instructions, subjects are presented with a table showing how many of the 20 subjects *preferred* Green (but not how many chose Green) in each period of the six previously conducted sessions of the baseline condition *TT-43*. We provided subjects with information observed in previous sessions so that they would observe that due to the large group size there is only a small variance in terms of how many subjects one should expect to switch preferences over time. Providing information about the number of people who on average/in expectation should switch type by a certain period could not convey this information. Thus, in condition *Public Awareness*, subjects should have *common* knowledge about the pace of preference change.

Preference Poll: Implements the same setting as in *TT-43* except that, at the start of period 14, subjects are asked what color they would prefer people in their group chose in the next periods. The individual (anonymous) responses at this poll are revealed. Then, after learning how many people answered Blue and how many Green, all subjects make their color choice for period 14. All aspects of the poll are explained in the experimental instructions. Subjects are thus aware at the start of the game that there will be a poll. In period 14, in expectation 75% of the subjects prefer Green and the probability that the group will have a majority preferring Green is 98.5%. The poll has two functions: aggregating preferences and providing a natural coordination point regarding mutual expectations about when to instigate change.

Incentive for Instigators: Implements the same setting as in *TT-43* except that subjects have an additional incentive to act as instigators of change. A reward is received by the four subjects who have persisted the longest in choosing the "majority color". The "majority color" is defined as the color chosen by more than 50% of the subjects in the final period (period 31). The reward of these "top four" subjects is that their earnings are raised to the level of the highest-earning subject in the session. If in period 31 each color is chosen by 10 subjects, no rewards are distributed. Initiating change to Green thus promises a reward, in particular, the costs incurred from leading change are made up for by the reward, but trying to instigate change is still risky, because there is no reward in case change fails to occur.

2. Elicitation of Normative Expectations in Separate Incentivized Survey

For a behavioral pattern to constitute a social norm, a critical condition is that most people believe that others *ought to conform* to it. In our experiment, do people think that most others believe they ought to choose Blue in period 1? How does the answer change over time as individuals' preferences gradually change? To address these questions, we ran an incentivized survey on Prolific, recruiting subjects with a similar socio-economic background as the subjects in our main experiment. The survey involved the following steps:

(1) Each participant had to read a condensed version of our experimental instructions from the main study. Individuals were allowed to complete the survey only if they correctly answered a set of comprehension questions. We collected 50 completed surveys.

(2) We presented each participant with four different scenarios. Each scenario corresponds to a situation that occurred in an actual session of the main study. Scenario 1 had two questions. First, we asked *"In your opinion, what is the socially most appropriate action for an individual to take in round 1?"* We explained that by "socially most appropriate", we mean the action that is "right" or most "ethical". The available answers were Blue, Green, and Both Equally Socially Appropriate. This question, therefore, elicits an individual's normative views (*2*). We found that 94% of the individuals answered that Blue is socially most appropriate; 6% answered that both colors are equally socially appropriate. Then, we asked *"What do you think is the color that most respondents in our survey consider to be socially most appropriate for an individual to choose in period 1?"* The second question was incentivized. Specifically, if an individual's response to the second question coincided with that given by the majority of respondents in the first question, a participant received $18.50 if Scenario 1 was selected for payment. This question elicits an individual's normative expectations (*2*). We found that 98% of the individuals answered Blue and 2% answered that both colors are equally socially appropriate. This illustrates that condition (*ii*) of the definition of a social norm is satisfied at the start of the experiment: people are aware that most other individuals believe that Blue is the "right" or most "ethical" choice in period 1.

(3) Subsequently, the participants were presented with scenarios 2, 3 and 4 (individuals received no feedback about the previous scenarios). These scenarios were similar to scenario 1, except that participants had to evaluate how socially appropriate the different actions are for periods 11, 21, and 31, respectively. One of the four scenarios was randomly selected for payment. Importantly,

when evaluating actions in periods 11, 21, and 31, the survey participants could observe the actual choices the subjects in the main study made up to this point. For example, in period 11, the survey participants could see the number of individuals in the main experiment who chose Blue or Green in periods 1 to 10. We chose a session of treatment *TT-30* for this. In this treatment, social tipping was observed in every society in the main experiment, allowing us to observe the change in normative beliefs as a society abandons an old social norm.

Figure S1 presents the results of the survey.

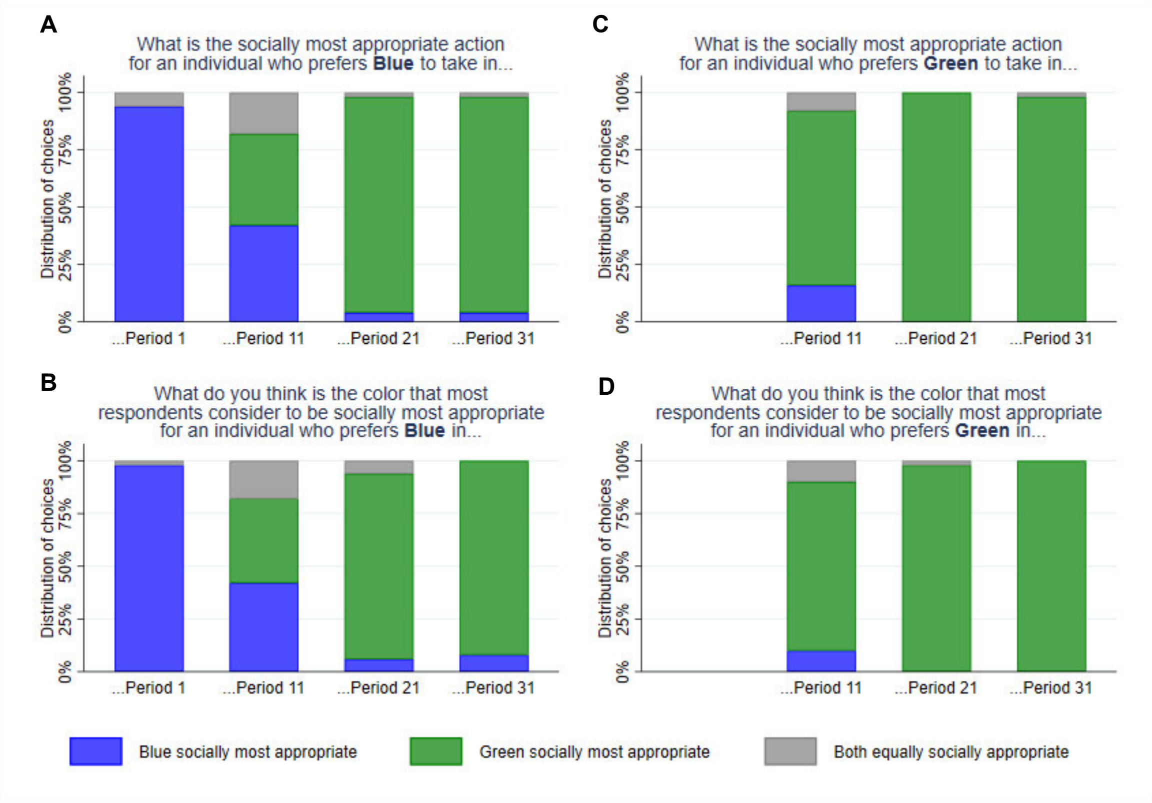

Fig. S1. Normative views and normative expectations over time. A) Normative views about actions by individuals who prefer Blue for periods 1, 11, 21 and 31. We find that 94% of the survey participants consider Blue to be the right/most ethical/socially most appropriate action in period 1. In period 31, 94% consider Green to be the right/most ethical/socially most appropriate action. **B)** Normative expectations about what others consider to be the socially most appropriate action for an individual who prefers Blue are mostly correct: the distribution of answers in figure B is similar to the one in figure A. Specifically, 98% of the survey participants correctly state that most others consider Blue to be the socially most appropriate action in period 1. In period 31, 92% correctly state that most others consider Green to be socially most appropriate. **C)** Normative views about actions by individuals who prefer Green in periods 11, 21 and 31 (in period 1, all individuals prefer Blue). Already in period 11, 76% of the survey participants consider Green to be socially most appropriate, and the percentage increases to 100% in period 21 and 98% in period 31. **D)** Normative expectations about socially most appropriate behavior for individuals who prefer Green yield a similar distribution of answers as in C. Taken together, figures A to D illustrate the existence of normative expectations in favor of Blue at the start of the experiment and in favor of Green at the end of the experiment. This indicates that Blue satisfies the conditions of a social norm in period 1, while Green satisfies the conditions in period 31, given that social tipping occurred. In Period 11, normative expectations are not aligned between subjects, indicating a phase of norm change.

3. Computation of Switching Thresholds and Model Simulation

In our model, similar to (*2-6*), each individual i is characterized by a different threshold f_i, which corresponds to the proportion of others who need to deviate from Blue to Green before individual i switches behavior from Blue to Green as well. A novelty of our model is that we provide a natural way of deriving these switching thresholds, reflecting the parameters of the environment. We explain this below.

Figure 1 in the article shows the payoff matrices pertaining to the social tipping game used in the study. Specifically, one can see that for an individual who prefers Green, the pecuniary payoff when choosing Green in a given period is $\pi(Green) = v_H - \boldsymbol{I}_{Miscoordination}\, p * (1 - g_t)$ and the pecuniary payoff when choosing Blue is $\pi(Blue) = v_L - \boldsymbol{I}_{Miscoordination}\, p * g_t$. The parameter v_H measures the benefit for choosing the (induced) preferred color (Green) and v_L measures the benefit for choosing the (induced) less preferred color (Blue). The indicator function $\boldsymbol{I}_{Miscoordination}$ equals 1 if two individuals fail to coordinate and 0 otherwise, i.e., the penalty applies only in case of miscoordination.

In addition to the monetized payoffs, an individual's willingness to deviate from the status quo is assumed to be affected by heterogeneity in personality traits/preferences as well as beliefs/expectations about their ability to expedite norm change. This heterogeneity is captured by the variable $\gamma_i \sim N(\mu, \sigma)$. That is, γ_i measures how individual i weighs the marginal benefit from change, $v \equiv v_H - v_L$, against the miscoordination penalty, p. For example, an individual who dislikes conformity would weigh the benefits higher than another individual who is happy to conform, relative to the miscoordination penalty. Similarly, an individual who expects that her deviation from the norm will accelerate change (thus receiving the marginal benefit from change, v, earlier) will also be characterized by a higher γ_i. Therefore, for an individual preferring Green, the perceived utility for choosing Green is $\pi(Green) + \gamma_i v$, while the perceived utility for choosing Blue is $\pi(Blue)$. Because γ_i models naturally-occurring heterogeneity between individuals, γ_i is not monetized in the experiment.

We note that an alternative way of writing an individual's perceived utilities is to treat γ_i as a preference shock such that the perceived utility for choosing Green is $(1 + \gamma_i) * v_H - \boldsymbol{I}_{Miscoordination}\, p * (1 - g_t)$ and the perceived utility for choosing Blue is $(1 + \gamma_i) * v_L - \boldsymbol{I}_{Miscoordination}\, p * g_t$. This way of writing the perceived utilities is mathematically equivalent to the one discussed above in our model, because the *difference* in perceived utilities for Green and Blue is the same.

We obtain for each individual i a switching threshold by equating the expected utilities for choosing Blue and Green (which depends on the expected color choice of the matched individual). The expected utility for choosing Green is $EU_i(Green) = v_H - p * \left(1 - g_t\right)^2 + \gamma_i v$. The expected utility for choosing Blue is $EU_i(Blue) = v_L - p * g_t^2$. The squared terms are a consequence of bilateral matching. To see this, note that for an individual who chooses Green, the probability of failing to coordinate with the matched individual is $1 - g_t$, as the latter is the fraction of other individuals choosing Blue and individuals are matched at random. The incurred penalty conditional on miscoordination is $p * (1 - g_t)$. Hence, the expected cost of miscoordination when choosing Green is $p * (1 - g_t)^2$. Analogously, the expected cost of miscoordination for someone choosing Blue is $p * g_t^2$. The switching threshold of individual i corresponds to the lowest value of g_t such that the expected utility for choosing Green is larger than the expected utility for choosing Blue, i.e., we can solve $EU_i(Green) = EU_i(Blue)$ for g_t. We obtain $f_i = 0.5 - 0.5\left(1 + \gamma_i\right) v/p$, where $v \equiv v_H - v_L$.

We define the *social tipping threshold,* denoted by f_{TT}, as the fraction of individuals who need to abandon the norm such that even individuals with $\gamma_i = 0$ have an incentive to follow and abandon the norm. Note that $\gamma_i = 0$ characterizes individuals who have no personal preference for change (e.g., nonconformity preferences) and who do not expect to expedite change by deviating from the norm. This implies that $f_{TT} = 0.5 - 0.5\, v/p$. Put differently, once the proportion of individuals choosing Green has reached f_{TT}, we expect change to be self-enforcing, as even individuals with $\gamma_i = 0$ want to abandon the Blue behavior. We can also express individual switching thresholds in terms of the social tipping threshold, in particular, $f_i = f_{TT} - 0.5\gamma_i v/p$.

Given a distribution of switching thresholds f_i and the rules describing the dynamics of change in the threshold model, one can simulate the proportion of individuals abandoning the initially established norm. Specifically, a single trial in our simulations involves three steps: (*i*) for each individual, we determine whether s/he prefers Blue or Green in the last five periods (consistent with Fig. 4 in the article), (*ii*) for each individual i, we draw γ_i from the probability distribution $N(\mu, \sigma)$ to compute the switching thresholds given by $f_i = f_{TT} - 0.5\gamma_i v/p$, where f_{TT}, v, and p follow directly from the treatment parameters, and (*iii*) the process of change is simulated based on the following rule: if g_t is the proportion of individuals who are believed to have abandoned the norm at the end of period t, then in period $t + 1$, all individuals with a threshold $f_i \leq g_t$ abandon the norm as well. For each such trial, we record the rate of norm abandonment, i.e., the fraction of individuals who choose Green when the process is completed. We ran 10,000 trials for each level of the tipping threshold. For any given tipping threshold, the mean over these trials is the probability of norm abandonment.

Robustness to Matching Procedure

The computation of the tipping threshold is robust to the way individuals are matched. The pairwise matching protocol we use is common in the literature (*7-9*), although other studies feature "group-wide" matching (*10-12*). In the case of group-wide matching, the expected utility for choosing Green is given by $EU_i(Green) = v_H - p * (1 - g_t) + \gamma_i v$ and the expected utility for Blue is $EU_i(Blue) = v_L - p * g_t$. Note that the squared terms disappear compared to the pairwise matching protocol, because group-wide matching implies that players always incur miscoordination costs if some players in their group choose the opposite color. However, the value of g_t for which $EU_i(Green) = EU_i(Blue)$, i.e., the individual threshold, are still given by the same expression, and in particular $f_{TT} = 0.5 - 0.5\ v/p$. Hence, our model can be used to analyze different matching environments. More generally, the idea of defining the social tipping threshold as the point of indifference for a fully myopic individual applies independently of the matching procedure. Note that the matching procedure may affect subjects' expectations about the prospects of change and thus the distribution of γ_i may change.

Estimation of Model Parameters and Switching Thresholds

Here, we describe the estimation technique we use to calibrate our model and to generate Figure 4. Following our model, the probability of an individual deviating from the established norm is given by $P(choice_t = Green) = P(f_i \leq g_{t-1})$. That is, an individual abandons the norm in period t if and only if her individual switching threshold f_i is below the fraction of others who have previously abandoned the norm. Plugging in $f_i = f_{TT} - 0.5\gamma_i v/p$, we obtain $P(choice_t = Green) = P(f_{TT} - 0.5\gamma_i v/p \leq g_{t-1})$ which after rearranging terms equals $P(\frac{f_{TT}-g_{t-1}}{0.5\ v/p} \leq \gamma_i)$. Letting $\tilde{\gamma} \equiv \frac{f_{TT}-g_{t-1}}{0.5\ v/p}$ and noting that $\gamma_i \sim N(\mu, \sigma)$, we obtain $P\left(\frac{\tilde{\gamma}-\mu}{\sigma} \leq z\right) = P\left(z < \frac{\mu-\tilde{\gamma}}{\sigma}\right) = \Phi(\frac{\mu}{\sigma} - \frac{1}{\sigma}\tilde{\gamma})$. This corresponds to a Probit model, where the estimated coefficient of the intercept provides an estimate of $\frac{\mu}{\sigma}$ and the coefficient of $\tilde{\gamma}$ provides an estimate of $-\frac{1}{\sigma}$. Hence, if multiplying the coefficient of $\tilde{\gamma}$ by -1 and taking the inverse we obtain an estimate for σ. Similarly, if dividing the coefficient of the intercept by the slope coefficient and multiplying the result by -1, we recover an estimate for μ (i.e., $-\frac{1}{\sigma} / \frac{\mu}{\sigma} *$

$(-1) = \mu)$. The standard errors of the estimates are derived using the delta method (nlcom command in the software package Stata). See the separate file "Data Analysis".

This results in an estimated distribution of $\gamma_i \sim N(1.73, 1.91)$. One interpretation of this results is that the average subject expects to expedite change by 1.73 periods when deviating from the norm. This is in line with the range of values we anticipated in Fig. 2A. Figure S2 below shows the distribution of switching thresholds implied by the estimated distribution of γ_i. For *TT-43*, almost all individuals have a switching threshold greater than 0 and the average individual switching threshold is around 35%. Thus, consistent with the data, social tipping is unlikely to occur. For *TT-43* with $\gamma_i \sim N(2.73, 2.25)$, which corresponds to the upper bound of the 99% confidence interval of the parameter estimates, we observe a leftward shift of the distribution of switching thresholds. The shift is small, however, confirming that small changes in expectations will not drastically alter the model's predictions. In contrast, for *TT-30*, with a lower tipping threshold and higher benefits of change (v is increased from 10 ECUs to 30 ECUs in this condition), more than 50% of the individuals are expected to be willing to instigate change (i.e., they have a negative switching threshold). Consistent with the data, change in *TT-30* is therefore likely to occur.

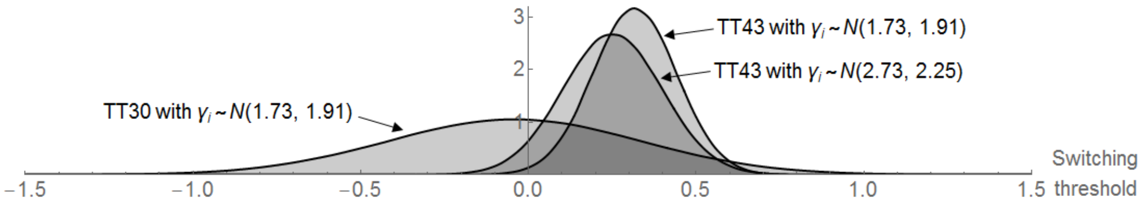

Fig. S2. Distribution of switching thresholds implied by model estimates. The estimated $\gamma_i \sim N(1.73, 1.91)$ for *TT-43* implies that almost all individuals have a switching threshold greater than 0 and the average threshold is around 35%. Further, even with $\gamma_i \sim N(2.73, 2.25)$, which corresponds to the estimated upper bound of the 99% confidence interval, we see that most switching thresholds still clearly exceed 0. Social tipping is thus unlikely to occur. In contrast, in condition *TT-30* with $\gamma_i \sim N(1.73, 1.91)$, there is a substantial leftward shift in the distribution of the switching thresholds: change in *TT-30* is thus likely to occur.

Discussion of Distributional Assumptions

In line with (*2*) and (*4*), we assume that $\gamma_i \sim N(\mu, \sigma)$. This assumption implies that γ_i can be negative for some individuals, raising the following question: what happens if we restrict γ_i to be positive? In addition, assuming $\gamma_i \sim N(\mu, \sigma)$ implies that individual switching thresholds, f_i, are also normally distributed and can fall outside the interval [0,1]. A second question that arises, therefore, is the following: how should f_i be interpreted?

With regards to the second question, rather than interpreting f_i directly as an individual's switching threshold, one can think of f_i as a latent variable measuring the willingness of switching to Green. A negative f_i indicates that individual i would be willing to be the first to deviate to Green even if the marginal benefit from change (v) was lower or the miscoordination penalty (p) was higher. In the data we only observe whether or not an individual instigates change. We do not know how far below 0 the latent variable f_i might be. Observed individual thresholds are thus the censored variables (at 0 and 1) of f_i. More intuitively, one could imagine that there are three types of individuals: those who are committed to change ($f_i \leq 0$), those who are committed to the status quo ($f_i \geq 1$), and individuals whose decision depends on the social dynamics observed over the periods of the game ($0 < f_i < 1$). This interpretation is in line with the literature on committed minorities (*8*), the difference being that in our experiment committed types emerge endogenously.

To address the first question, a negative γ_i indicates that individual i chooses Blue even when the society the individual belongs to reaches the tipping threshold. A negative γ_i is thus best interpreted as a status quo bias due to preferences for the established norm (beyond the induced pecuniary incentives) or pessimistic expectations about the cost involved when transitioning to Green. Individuals with a negative γ_i should be a minority, as the most natural assumption is that people consider, to some extent, that their deviation to Green may motivate others to deviate as well. In line with this, the distribution estimated from the data, $\gamma_i \sim N(1.73, 1.91)$, implies that 18% of the individuals have a negative γ_i.

However, one might argue that γ_i should lie in the interval [0,∞], especially when interpreting γ_i as an expectation about the future benefits from instigating change today. The exponential distribution offers a plausible case for this scenario: its support is [0,∞] and smaller values of γ_i are more likely than larger values. Figure S3 shows the predictions for the exponential distribution and compares them with the predictions for the normal distribution, holding constant the mean of γ_i. The predictions are qualitatively similar. This finding may not hold for other distributions, in particular multimodal distributions that are more prone to situations where social tipping is instigated but fails to fully reverse the social equilibrium. We leave this as an interesting question for future research.

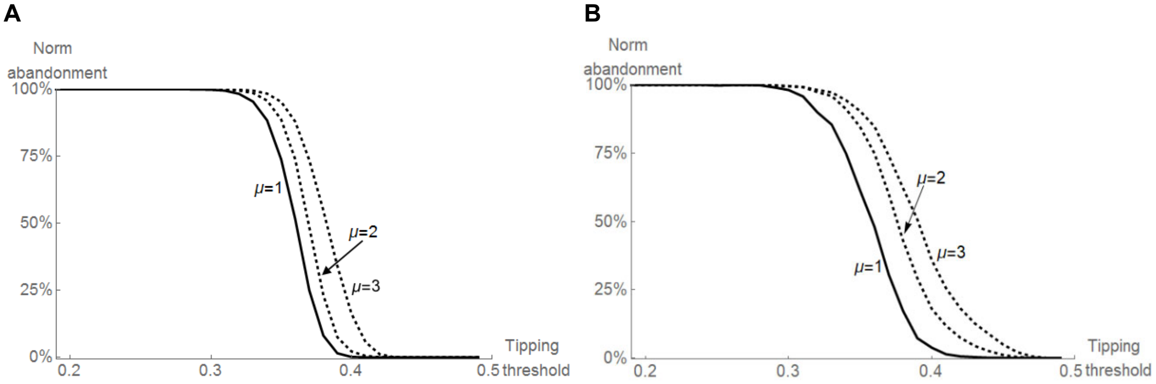

Fig. S3. Theoretically predicted norm abandonment for normal and exponential distribution. A) Probability of norm abandonment for different tipping thresholds assuming $\gamma_i \sim N(\mu, 1)$. **B)** Probability of norm abandonment for different tipping thresholds assuming $\gamma_i \sim Exp(1/\mu)$, which implies the same mean as in A. For both distributions, the predictions show that the probability of norm abandonment decreases rapidly above a tipping threshold of 35%. The decrease is slower for the exponential distribution, i.e., there is a larger range of tipping thresholds for which the probability of norm abandonment is strictly between 0% and 100%.

4. Additional Data Analysis

Penalty Choices in Condition *TT-Endo*

Figure S4 displays the penalty choices of subjects choosing Blue (i.e., the penalties faced by subjects deviating from the norm to choose Green) in condition *TT-Endo*. As can be seen, the proportion of subjects choosing a high miscoordination penalty of 7 ECUs per subject choosing the other color is increasing over time, and as a result, the average penalty is also increasing. The increase in average penalties is significant (P<.001, linear random effects model regressing the penalty choice on time). Interestingly, the penalty choice does not significantly differ between types, i.e., whether a subject prefers Blue or Green. This suggests that independent of their preferences subjects increase the sanctions for norm violators over time to avoid miscoordination costs. We also find that subjects who have incurred high penalties in previous periods are more likely to choose high miscoordination penalties in the current period (P=.017, linear random effects model regressing the penalty choice on average penalty and incurred penalty two periods ago, which is the last period for which others' behavior is observed). This could be due to indirect retaliation or due to an increased urgency for signaling that deviations from the norm should be avoided.

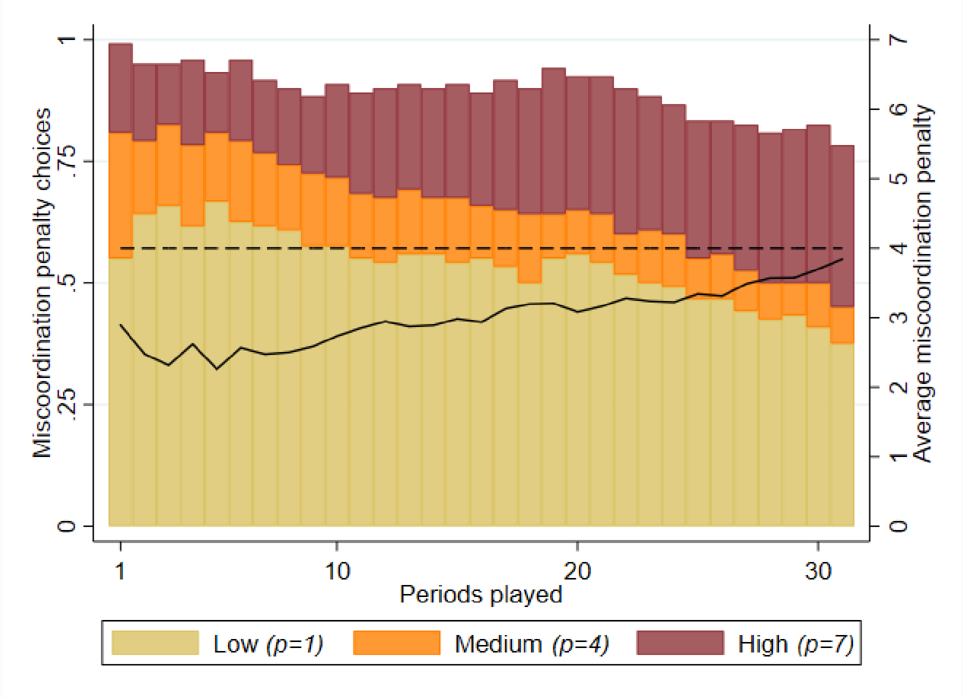

Fig. S4. Miscoordination penalty choices over time. Penalty choices of subjects choosing Blue (i.e., penalties faced by subjects deviating to Green) in *TT-Endo*. Here, the penalty, p, refers to the cost a subject who fails to coordinate incurs *per subject* in the group choosing the other color. The total height of each bar gives the fraction of subjects choosing Blue. Each bar is composed of three regions, the fraction of subjects choosing a low (bottom part), medium (middle part), and high penalty (top part). The fraction of subjects choosing a high penalty is increasing over time, leading to an increase in the average penalty (solid line), and hence an increase over time in the pressure to conform.

Out-of-sample Predictions

Figure 4 in the article shows the predictions of the estimated model and the 99% confidence interval for the conditions that vary the tipping threshold (*TT-43*, *TT-30*, *TT-23*, and *TT-Endo*). This provides an in-sample test of our theoretical model. Here, we also provide out-of-sample tests. Specifically, we estimate the model using only two of the above four conditions and test whether the model predictions are in line with the behavior observed in the other two conditions. The results displayed in Fig. S5 provide an affirmative answer – focus in particular the point predictions given by the solid lines as, naturally, the 99% confidence intervals are less precise than in Fig. 4 where we estimate the model using all data.

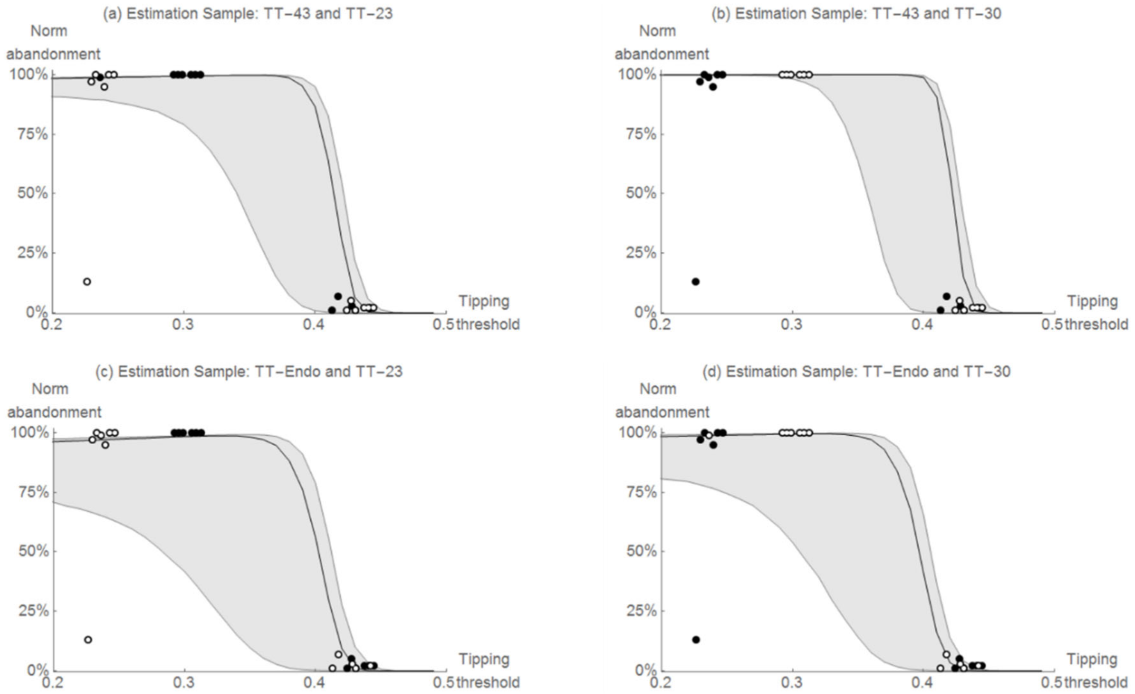

Fig. S5. Norm abandonment as a function of the tipping threshold with out-of-sample predictions. Each marker represents the percentage of subjects in the last five periods that abandoned the Blue norm for a given experimental society. Filled markers represent the out-of-sample observations that we aim to predict; unfilled markers represent the observations included in the estimation sample. The theoretically predicted frequency of norm abandonment (solid line) and the 99% confidence interval (shaded area) are averages from 10,000 simulated trials per tipping threshold based on the estimated parameters (Probit model with society random effects). The theoretical predictions correctly anticipate norm abandonment in most societies.

Time Series of Experimental Data

Figure S6 shows time series of behavior over the 31 periods in all 54 societies. The numerical data set is provided in a separate file “Dataset S1”.

The left panel in Fig. S6 displays behavior over time in the conditions that vary the tipping threshold, which are discussed extensively in the article. The right panel displays, for a constant tipping threshold of 43%, the effect of the different conditions affecting subjects’ expectations for change (and for condition *Incentive for Instigators*). For these conditions, it is instructive to derive the implied increase in γ_i relative to the baseline estimate of $\mu = 1.73$. To do so, we determine the value of μ that is consistent with the observed behavior in *Fast Feedback*, *Small Society*, *Public Awareness*, and *Preference Poll*, holding constant the variability at the baseline estimate of $\sigma = 1.91$. We find that the mean beliefs that rationalize observed behavior are $\mu = 2.4$ for *Fast Feedback* (a 39% increase relative to $\mu = 1.73$), $\mu = 5$ for *Small Society* (a 189% increase relative to $\mu = 1.73$), $\mu = 6.2$ for *Public Awareness* (a 258% increase relative to $\mu = 1.73$), and $\mu = 7.7$ for *Preference Poll* (a 345% increase relative to $\mu = 1.73$). Our model thus provides a way of measuring the increase in optimism in the conditions designed to affect expectations for change. In particular, conditions that alter *collective* expectations (*Public Awareness*, *Preference Poll*) lead to a more than twofold increase in individuals’ expectations about their ability to successfully instigate change.

It is also noteworthy that in *Fast Feedback* the probability of instigating change (choosing Green when the tipping threshold has not been reached) is significantly lower than in *TT-43* (P<.001, random effects Probit regression with society-cluster standard errors). Finally, condition *Small Society* leads to the lowest payoffs of all nine conditions, mostly due to the high miscoordination costs associated with the slow transitioning from Blue to Green in the cases where change occurred (see also *Fig. S7*).

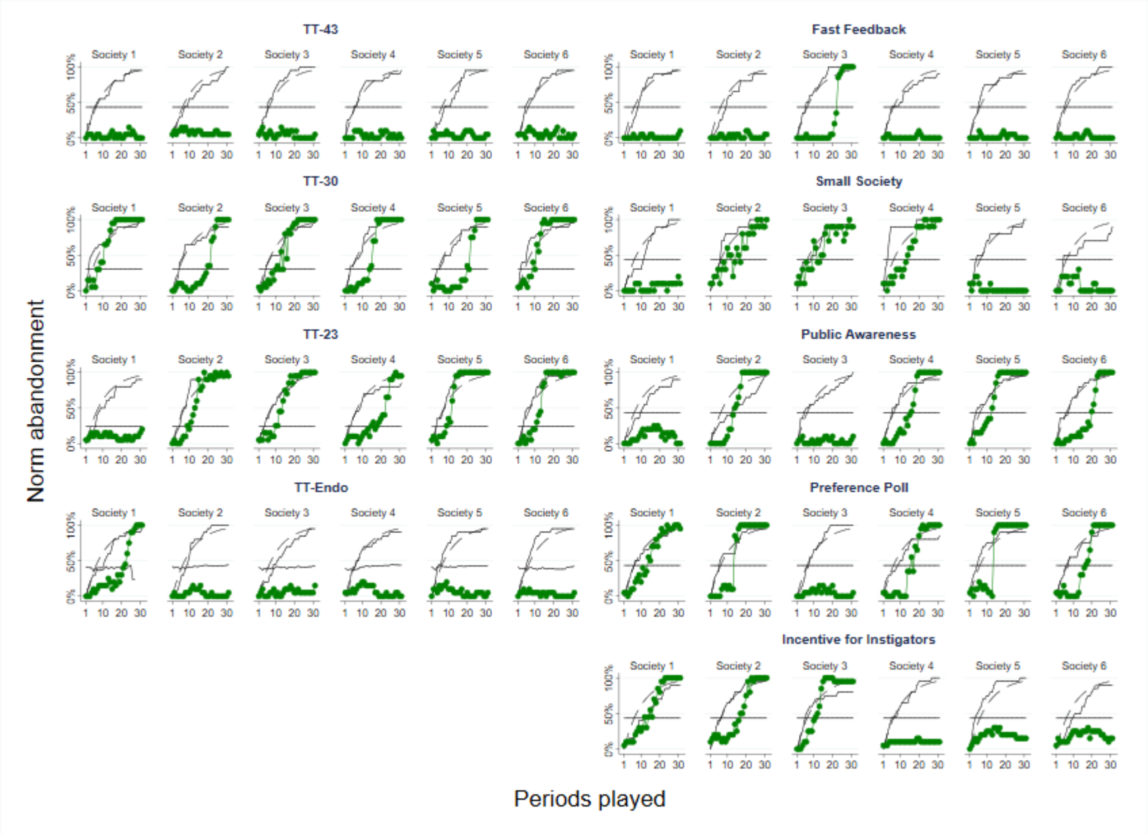

Fig. S6. Time series of norm abandonment for all experimental conditions and societies. Norm abandonment is shown as the line with circled markers. The tipping threshold is given by the horizontal line. The dashed concave line indicates the theoretically expected fraction of subjects preferring to abandon the norm; the solid increasing line the corresponding realized fraction. The column on the left shows that the tipping threshold is a crucial determinant of the probability of social tipping and that, if given the opportunity to lower social penalties (*TT-Endo*), societies fail to do so and are trapped at the detrimental “Blue norm”. The column on the right displays, for a constant tipping threshold of 43%, the effect of different conditions affecting subjects’ expectations for change and of a condition providing incentives for subjects who successfully instigate social tipping.

Payoff Loss Relative to First-Best Outcome

Figure S7 shows the loss in average payoffs (efficiency) relative to the first-best outcome, which is achieved when the entire society changes behavior from Blue to Green in the first period in which the majority of subjects prefers Green (except in *TT-30*, where the change should occur earlier due to the larger benefits from change, v). Payoff losses compared to the socially optimal outcome can be due to abiding to the detrimental (inefficient) norm or due to penalties from miscoordination. As Fig. S7 shows, both factors are important. Condition *TT-23*, where penalties are small, is the condition with the lowest payoff losses, less than 10% efficiency loss relative to the socially efficient outcome. Condition *Small Society* is the condition with the highest payoff losses, almost 40% efficiency loss relative to the socially efficient outcome, mainly due to miscoordination. Indeed, in *Small Society* the average miscoordination penalty incurred per subject and period is 8.85 (random effects regression with society-clustered standard errors), significantly higher than the corresponding penalty of 3.59 in *TT-43* (P=.014). The average miscoordination penalty incurred per subject and period in the other conditions is between 1.41 in *Fast Feedback* and 5.55 in *TT-30.* The only exception is condition *Incentive for Instigators* with a similar degree of miscoordination as in *Small Society* (P=.934), but there miscoordination penalties are partly offset by the external reward to lead change.

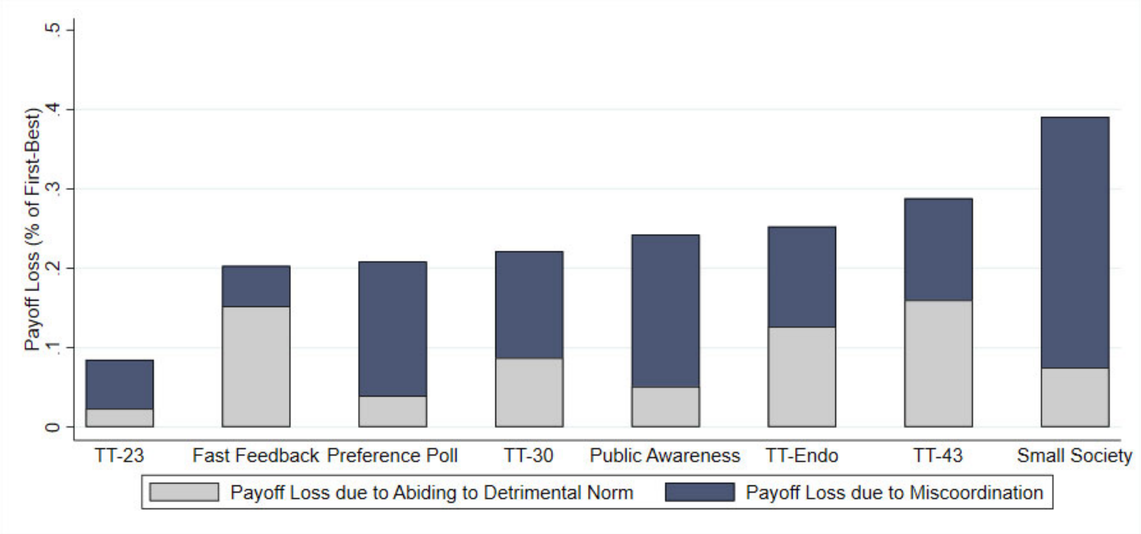

Fig. S7. Loss in average payoffs relative to first-best outcome. Payoff losses in percent relative to the potential payoffs in the socially efficient outcome (which for most conditions means switching from Blue to Green once a majority prefers Green) for the different experimental conditions. The lower part of each bar shows the payoff loss due to inefficient color choices, i.e., choosing Blue when Green would be socially efficient and vice versa. All conditions except *Fast Feedback* and *TT-Endo* outperform the baseline *TT-43* in terms of avoiding adherence to the detrimental norm (P<.004, random effects regressions with society-clustered standard errors). The upper part shows the payoff loss due to miscoordination penalties. *Small Society* exhibits the largest payoff losses due to miscoordination among all conditions (P=.014 compared with *TT-43,* random effects regressions with society-clustered standard errors).

Instigators of Change

Figures S8 and S9 provide additional analyses for instigators of change, i.e., for individuals who deviate from Blue before the social tipping threshold is reached. Specifically, Fig. S8 provides information about the cost of instigating change, and Fig. S9 discusses the consistency of observed behavior with the assumption of threshold models that individuals have unique switching threshold.

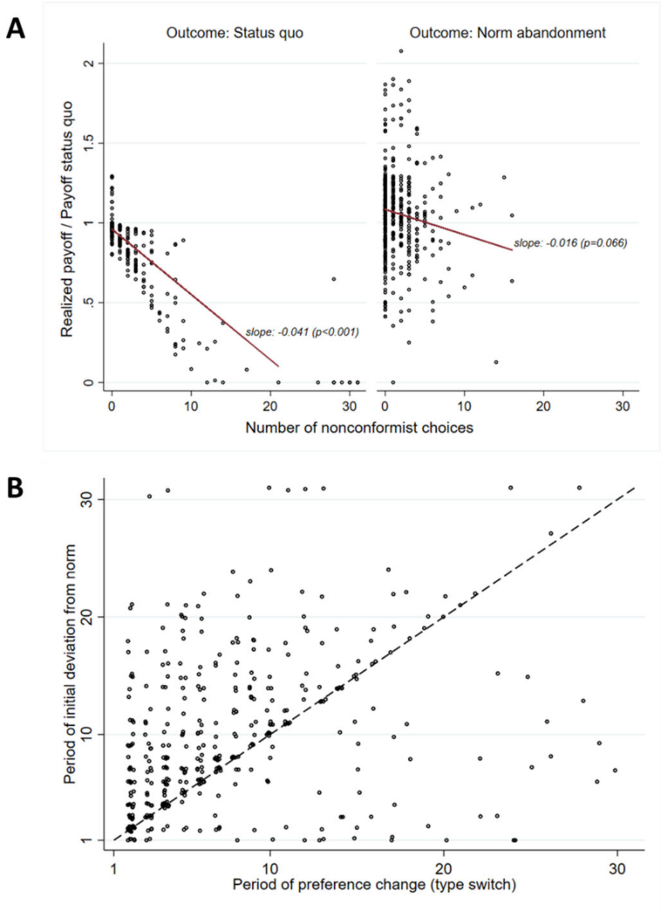

Fig. S8. Instigators of change. A) Realized earnings normalized by the earnings an individual would have made if everyone adheres to the norm in all periods plotted against the number of times an individual deviated from the norm (nonconformist choices). If the outcome in a society is that the status quo prevails, i.e., the less than 50% of individuals have abandoned the norm by the final period, each deviation from the norm on average causes a 4.1% points loss in normalized payoffs, and most individuals would be better off if no deviations had happened. The latter follows because the normalized payoff is below 1. Even if in a given society norm abandonment is successful, instigators of change typically have a normalized payoff below 1. This suggests that individuals are motivated to instigate norm abandonment despite the, on average, negative effect on their expected payoffs. **B)** Period of initial deviation from the norm plotted against the period in which an individual's preference changed. The clusters near the 45°-line show that many initial deviations occur in the same period as an individual's type changes, or shortly thereafter. On the other hand, many observations also lie substantially above the 45° line, which shows that individuals who prefer Green either strategically delayed their deviations to a later point in time when others are more likely to follow or waited for others to instigate change first.

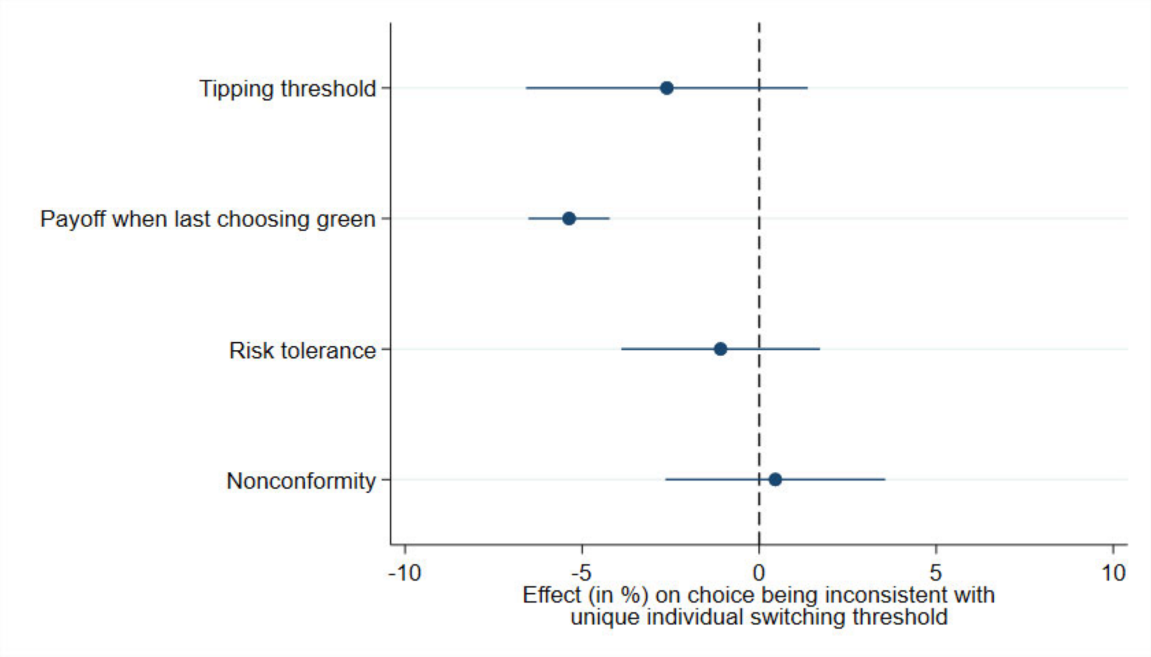

Fig. S9. Incurring high miscoordination penalties leads to choices that are inconsistent with a strict interpretation of the threshold model. Effects (in %) and 99% confidence intervals on the probability that a choice is inconsistent with a unique individual switching threshold (linear random effect model with society-clustered standard errors). Inconsistency with a unique individual switching threshold is defined in the caption of Fig. 6 in the article. The lower the payoff in the last previous period in which an individual chose Green, the more likely the individual is to make a choice that is inconsistent with a unique switching threshold. The effect is substantial: a payoff reduction of 76 ECUs, which corresponds to the miscoordination penalty faced by an instigator of change in *TT-43,* corresponds to a 41% increase in observing an inconsistent choice in a future period. In addition, 81.4% of all inconsistent choices occur when an individual switches back from choosing Green to choosing Blue. High penalties for failing to conform thus discourage instigators of change, who at least temporarily revert back to Blue.

5. Tipping Threshold and Committed Minorities

Several important studies in the literature rely on models that emphasize the existence of a minority of actors in a society that is committed to inducing change (*5, 8*). We can amend our model to account for such "committed minorities". To illustrate this, we apply our model to the setting of (*8*), who study committed minorities and social conventions using an agent-based model.

In the model of (*8*), the key parameter is an agent's memory length, determining the number of times an agent needs to be exposed to a different social convention before switching behavior as well. For making predictions, informed by their previous research (*13*), the authors assume that agents have a memory length of 12 periods. In contrast, in our threshold model, actors base their decisions on the proportion of others who have already abandoned a norm and, in addition, actors are *heterogeneous*, as they differ in their expectations about the likelihood of change. This allows us to study the emergence and characteristics of change instigators, in particular. See also (*14*) for a discussion of different approaches to modeling the emergence of social consensus.

In (*8*), players in each period earn a payoff of x (\$0.1) if they coordinate and lose the same amount if they fail to coordinate. Players' only concern is to coordinate. However, there is also a fraction of committed players, who always choose the alternative behavior. We denote this fraction by f_c. Using our terminology, the committed players always choose Green. In (*8*), the committed minority is introduced via confederates after the experimental subjects have reached a consensus, or an established convention. Using our terminology, this is the Blue convention.

In this setting, following our approach, the expected utility for an individual choosing Blue is given by $\mathrm{EU}_i(Blue) = (1 - g_t)x - (g_t + f_c)x - \gamma_i f_c x$. The first term corresponds to the probability of being matched with a player who chooses Blue multiplied with the benefit from coordinating. The second term captures the expected cost of failing to coordinate. The third term captures expectations for earning low benefits in the future due to miscoordination with the committed minority. The expected utility for choosing Green is given by $\mathrm{EU}_i(Green) = (g_t + f_c)x - (1 - g_t)\, x + \gamma_i f_c x$. The third term captures the expectation for earning a high benefit in the future due to coordination with the committed minority. The value of g_t for which $\mathrm{EU}_i(Blue) = \mathrm{EU}_i(Green)$, the individual switching threshold, is given by $f_i = f_{TT} - 0.5\gamma_i f_c$. The social tipping threshold when $\gamma_i = 0$ is given by $f_{TT} = 0.5 - f_c$. The tipping threshold takes a simple form, because in (*8*) the impetus for change is solely the committed minority.

Based on the individual thresholds, one can estimate the distribution of γ_i using the dynamics of the threshold model. Two remarks are in order. First, in (*8*) subjects are unaware of the introduction of a committed minority; in our setting the preference change is public knowledge. In other words, the circumstances that create a need for change are public knowledge only in our setting. Second, in (*8*) subjects learn about the fraction who have adopted a new convention via observing the choice of their matches over time but do not directly observe the proportion of individuals who have adopted a new convention in a given period. In our experimental environment, subjects are informed about the fraction of others that have chosen to abandon the established norm. Both remarks suggest that subjects' expectations about the likelihood of change are likely lower in (*8*) than in our experiment (i.e., we expect that the distribution of γ_i shifts to the left).

When estimating our model using the data from (*8*), we allow for the existence of a committed minority, and we assume that subjects' estimates about the proportion of others who have adopted the new convention is the average choice from their matches in the previous 12 periods (memory length). We obtain an estimated distribution of $\gamma_i \sim N(0.74, 1.26)$. This corresponds to a leftward shift in the distribution compared with the estimate for our experiment (where μ = 1.73, $\sigma = 1.91)$. Moreover, Fig. S10 shows that our model predicts behavior of the experimental societies in (*8*) well: all observations are within or at the boundary of the 99% confidence interval of the theoretical predictions. The accuracy of the threshold model at predicting tipping of social conventions in a different experimental setting suggests that it can be used to study societal change broadly.

Finally, it is interesting to note that in (*8*) change is not observed at a tipping threshold of around 30% (see Fig. S10), or rather for the size of the committed minority that corresponds to a 30% tipping threshold based on our transformation. In our setting – with public knowledge of the process of preference change and feedback about past behavior of the entire group – a tipping threshold of 30% (*TT-30*) resulted in complete norm abandonment in all six societies. This suggests that if we were to re-run our experiment but remove public knowledge about the preference change as well as feedback about past group behavior, as in (*8*), we would likely observe persistence of the detrimental norm, even at a tipping threshold of 30%.

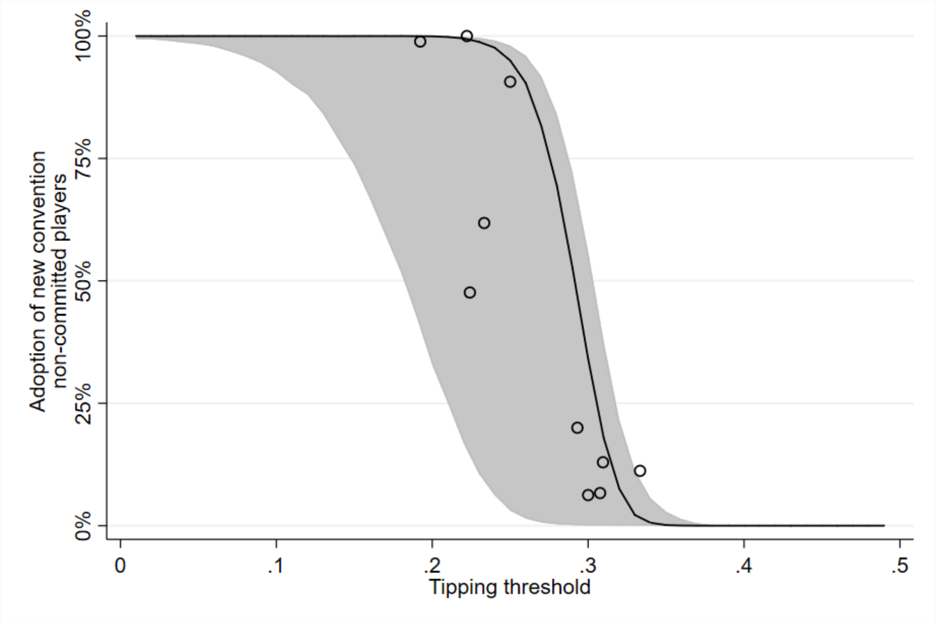

Fig. S10. Adoption of new convention for different tipping thresholds. The markers represent, for the ten experimental societies in (*8*), the percentage of choices by non-committed individuals in the last five periods that do not correspond to the initially established convention. Also shown are the theoretical predictions based on the parameters estimates $\mu = 0.74$ and $\sigma = 1.26$ (solid line). The shaded area shows the corresponding 99% confidence interval. The predictions from our model provide a good approximation of the empirical findings of (*8*). Interestingly, the parameter estimates are lower/less conducive to change for the data from (*8*) than for our data, demonstrating that expectations crucially depend on public knowledge of preferences.

6. Separate Files

Experimental Instructions S1. Instructions subjects received at the start of the experiment in the different conditions and the incentivized survey.

Dataset S1. Full data set for all 54 experimental sessions.

Data Analysis S1. Code used to analyze the data including all regressions. Allows replication of empirical figures.

Model Simulation S1. Code to simulate the threshold model. Allows replication of theoretical predictions.

SI References

1. R. Goldsmith, R. Clark, B. Lafferty, Tendency to conform: a new measure and its relationship to psychological reactance. *Psychological Reports* 96, 591-594 (2005).

2. C. Bicchieri, *Norms in the wild: how to diagnose, measure, and change social norms* (Oxford University Press, 2016).

3. T. Schelling, *Micromotives and macrobehavior* (WW Norton & Company, New York, 1978).

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

5. P. Oliver, G. Marwell, R. Teixeira, A theory of the critical mass: I. interdependence, group heterogeneity, and the production of collective action. *American Journal of Sociology* 91, 522-556 (1985).

6. M. Macy, Chains of cooperation: threshold effects in collective action. *American Sociological Review* 56, 730-747 (1991).

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

8. D. Centola *et al.*, Experimental evidence for tipping points in social convention. *Science* 360 1116-1119 (2018).

9. D. Acemoglu, M. Jackson, History, expectations, and leadership in the evolution of social norms. *The Review of Economic Studies* 82, 423-456 (2014).

10. W. Brock, S. Durlauf, Discrete choice with social interactions. *The Review of Economic Studies* 68, 235-260 (2001).

11. L. Blume, W. Brock, S. Durlauf, R. Jayaraman, Linear social interactions models. *Journal of Political Economy* 123, 444-496 (2015).

12. D. Smerdon, T. Offerman, U. Gneezy, 'Everybody's doing it': on the persistence of bad social norms. *Experimental Economics*, 1-29 (2019).

13. D. Centola, A. Baronchelli, The spontaneous emergence of conventions: an experimental study of cultural evolution. *Proceedings of the National Academy of Sciences* 112, 1989-1994 (2015).

14. A. Baronchelli, The emergence of consensus: a primer. *Royal Society Open Science* 5, 172189 (2018).