Appendix – Supplementary Information (SI):

Predicting social tipping and norm change in controlled experiments

James Andreoni[a], Nikos Nikiforakis[b,*], Simon Siegenthaler[c]

[a]Department of Economics, University of California, San Diego, 9500 Gilman Dr. #0508, La Jolla, CA 92093. [b]Division of Social Science, New York University Abu Dhabi, P.O. Box 129188, Abu Dhabi, United Arab Emirates, [c]Naveen Jindal School of Management, University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080

* 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

1. Materials and Methods: Additional Information

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). 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 only upon subjects' arrival at the lab. A total of 54 sessions was run with 1020 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 sheet.

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 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.

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.*

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 four of the 1020 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 change, (*ii*) people's expectations about the likelihood of change and their contribution to it (self-efficacy), and (*iii*) the presence of individuals who are willing to lead change. 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 alternatives: 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 30 experimental points and choosing the other color a payoff of 20 experimental points. Hence, the benefit from change is $b = 30 - 20 = 10$ points per period. In case subjects fail to choose the same color, they incur a miscoordination penalty of 4 for each subject in the group choosing the other color. The maximum penalty is $p = 76$ (19 subjects choosing the other color times a penalty of 4 per subject). These parameters imply a tipping threshold of 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 players in the group 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 to 50 experimental points. Hence, the benefit from change equals $b = 50 - 20 = 30$. The tipping threshold in this condition is at 30%, i.e., just below the theoretical cutoff for tipping (see *Fig. 1*).

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 points to 1 point. The maximum penalty is therefore $p = 19$ (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. 1*).

TT-Endo: Implements the same setting as in *TT-43* except that, in each period, subjects choose the miscoordination penalty their matched subject receives per subject in the group choosing the other color. The available choices are a miscoordination penalty of 1 as in *TT-23*, of 4 as in *TT-43*, or of 7 (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.

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. This keeps the maximum penalty p 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 points, which equals 76 points. In *Small Society* the total penalty is the same, 9 players times 8.44 points, which also equals 76 points. 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 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 by 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 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. Additional Data Analysis

Penalty Choices in Condition *TT-Endo*

Figure S1 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 points 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.

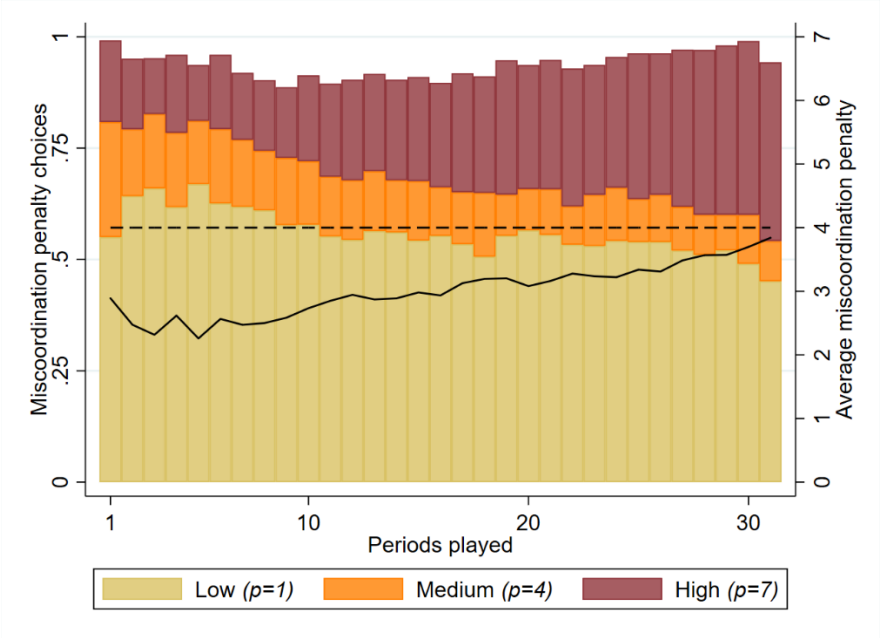

Fig. S1. Miscoordination penalty choices over time. Penalty choices of subjects choosing blue (i.e., penalties faced by subjects deviating to green) in *TT-Endo*. Here, the penalty refers to the cost a subject who fails to coordinate incurs *per subject* in the group choosing the other color. The total 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 3 in the article shows the predictions of the theoretical model and the 99% confidence interval based on the estimated parameter values from 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. S2 provide an affirmative answer – note in particular the point predictions given by the solid lines as, naturally, the 99% confidence intervals are less precise than in Fig. 3 where we estimate the model using all data.

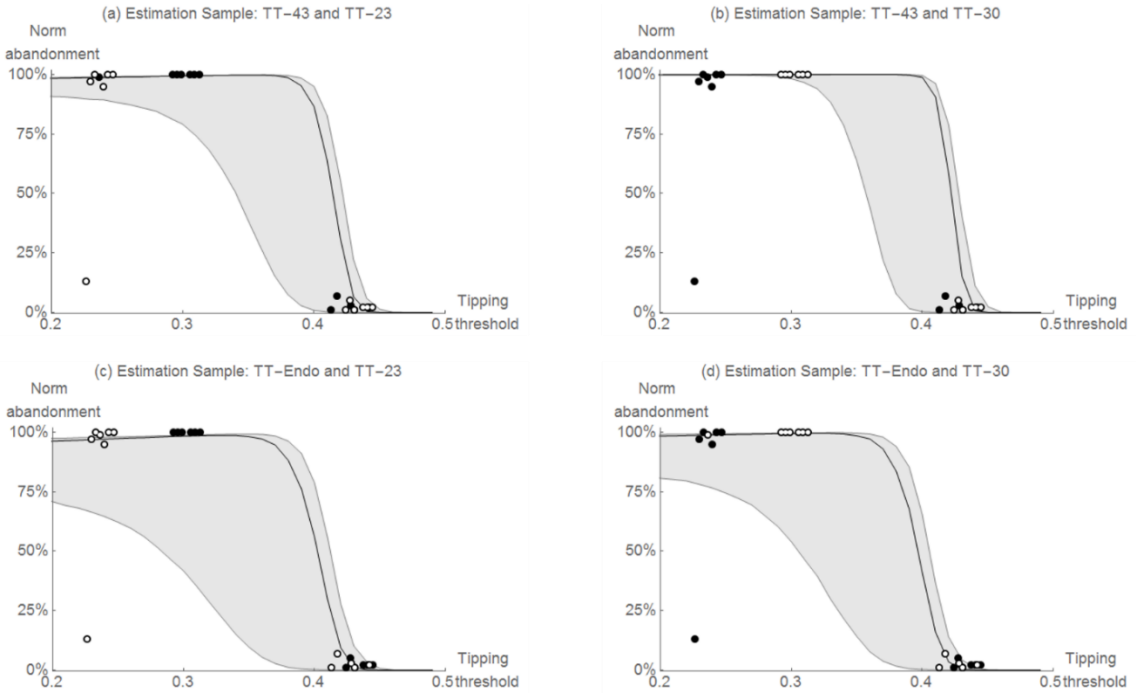

Fig. S2. Norm abandonment as a function of the tipping threshold with out-of-sample predictions. Each marker represents the percentage of subjects in the last five periods that abandoned the blue norm for a given experimental society. Filled markers represent the out-of-sample observations that we aim 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 S3 shows time series of behavior over the 31 periods in all 54 societies. The numerical data set is provided in a separate file "Data".

The left panel in Fig. S3 displays behavior over time in the conditions that vary the tipping threshold, which are discussed extensively in the article. The right panel in Fig. S3 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 direct 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 the benefits from instigating 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. S4*).

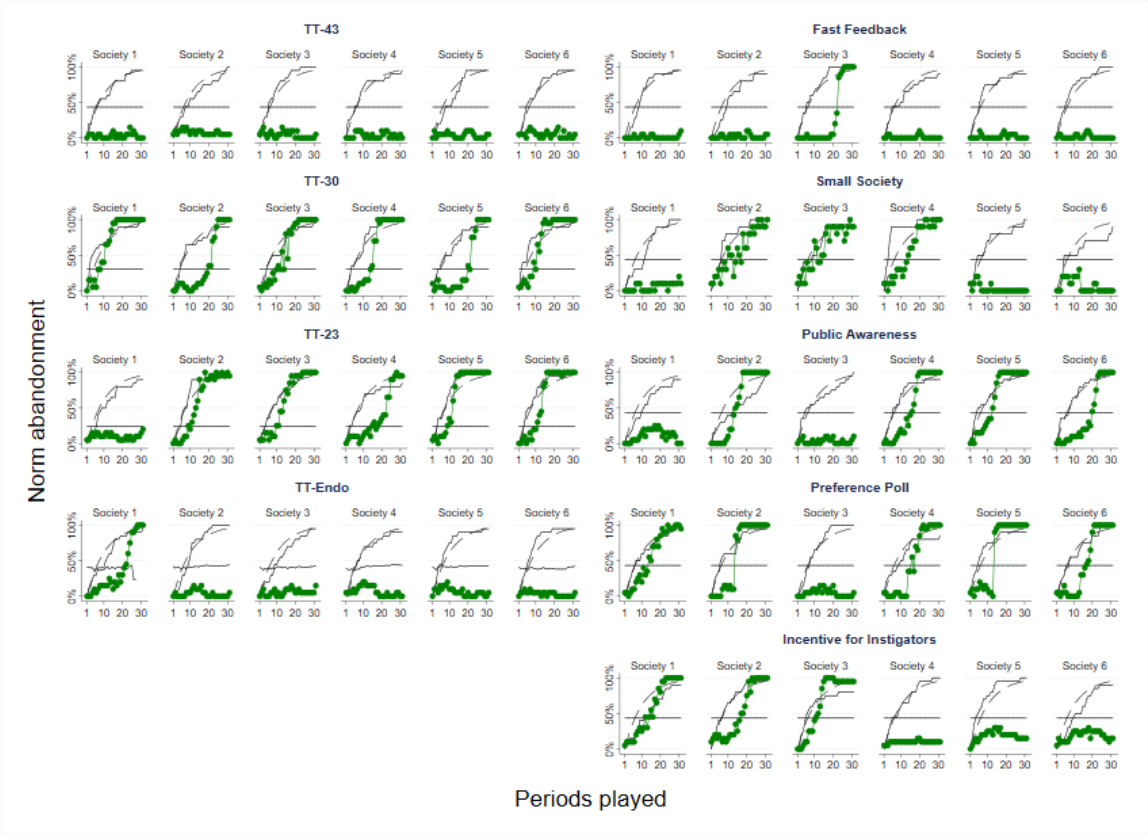

Fig. S3. Time series of norm abandonment for all experimental conditions and societies. Norm abandonment is shown as the line with circled markers. The tipping threshold is given by the horizontal line. The dashed concave line indicates the theoretically expected fraction of subjects preferring to abandon 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 S4 shows the loss in 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 b). 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. S4 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 the baseline *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 incentive to lead change.

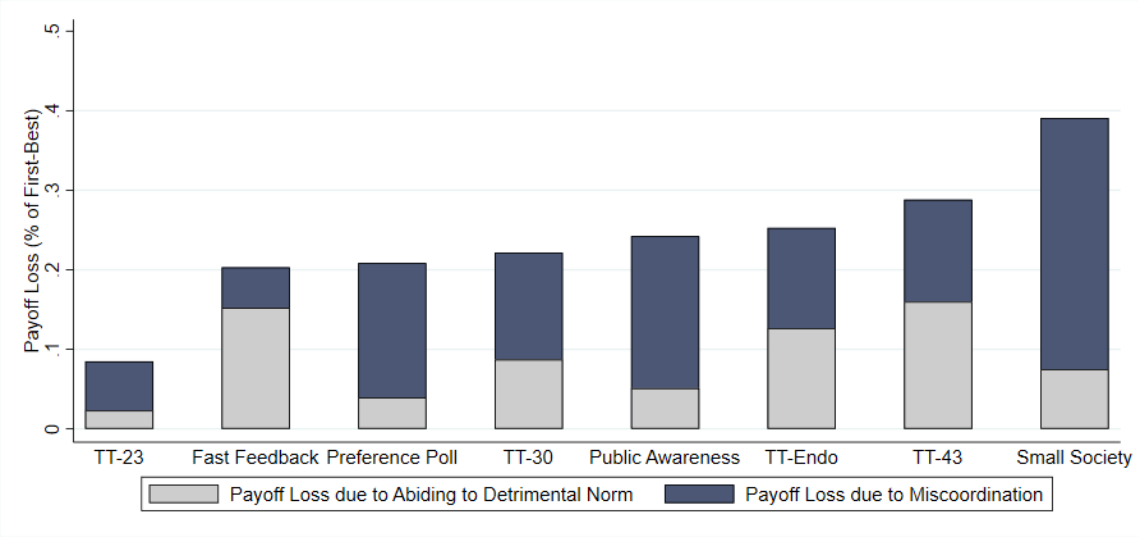

Fig. S4. Loss in total payoffs relative to first-best outcome. Payoff losses in percent relative to the potential payoffs in the socially efficient outcome (which for most conditions means switching from blue to green once a majority prefers green) for the different experimental conditions. The lower 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). Condition *Incentive for Instigators* is not shown as there instigators of change could recoup their earnings due to the exogenous incentive.

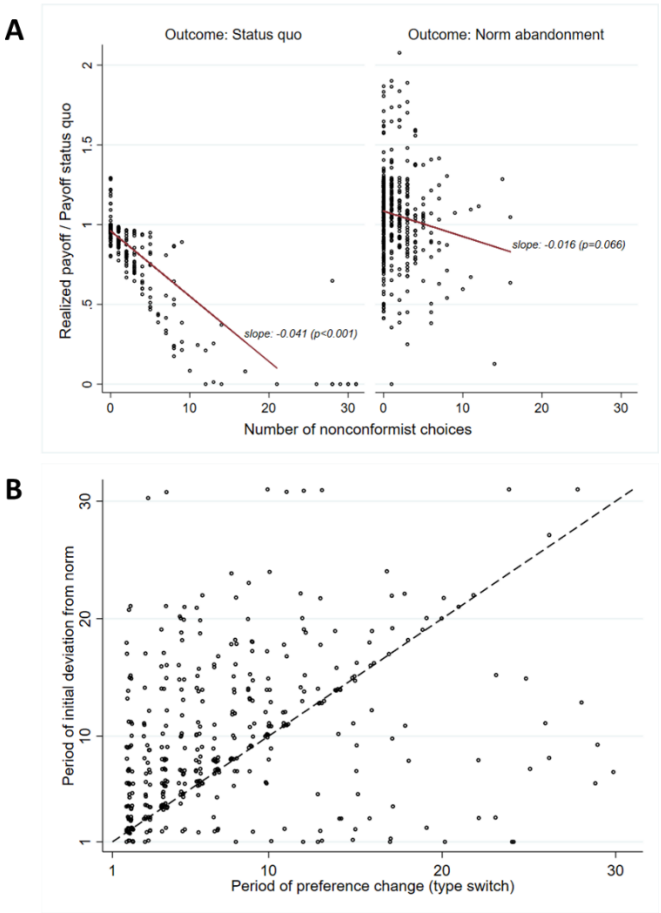

Fig. S5. Instigators of change. A) Realized earnings normalized by the earnings an individual would have made if everyone adheres to the norm in all periods plotted against the number of times an individual deviated from the norm (nonconformist choices). If the outcome in a society is that the status 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% 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 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 shows 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.

3. Computation of Thresholds and Model Simulation

In our model, similar to (*2-6*), each individual i is characterized by a different switching threshold f_i. A novelty of our model is that we provide a natural way for deriving the switching thresholds. In particular, we extend the traditional threshold model of (*3)* and *(4)* by providing a foundation for the different switching thresholds assuming that agents are rational decision-makers (who may have different and possibly wrong beliefs about the likelihood of change). This also allows us to pin down a social tipping threshold, denoted by f_{TT}. It corresponds to the fraction of group members that have already abandoned the norm such that the expected payoff from deviating from the norm starts to exceed the expected payoff from adhering to it, even for subjects that disregard potential future benefits from their own norm deviation. Once reached, f_{TT} is a point of no return beyond which norm abandonment occurs with a probability of 100%.

To derive individual switching thresholds f_i, we first compute the expected payoffs from choosing green and blue for an individual who has a preference for green. The expected payoff from choosing green is given by $\Pi_i(green) = b_{green} - f_{blue}^2 p + \gamma_i b$. The expected payoff from choosing blue is given by $\Pi(blue) = b_{blue} - f_{green}^2 p$. To understand why these expected payoffs correspond directly to the incentives subjects face in the experiment, note the following. The first term in both expressions is an individual's benefit from choosing green, denoted by b_{green}, and an individual's benefit from choosing blue is denoted by b_{blue}. The second term in each expression is the expected cost of miscoordinating. It is given by the probability of miscoordinating multiplied by the penalty incurred when this happens. For an individual who chooses green, the probability of miscoordinating is f_{blue}, the fraction of other individuals choosing blue. The incurred penalty is the fraction of other subjects choosing blue, f_{blue}, times the maximum penalty p. Hence, the expected cost of miscoordinating for someone choosing green is $f_{blue}^2 p$. Similarly, the expected cost of miscoordination for someone choosing blue is $f_{green}^2 p$. The last term in $\Pi_i(green)$ reflects the expected benefit from instigating/expediting change, which depends on self-efficacy γ_i and the net benefit of change $b \equiv b_{green} - b_{blue} > 0$. Because choosing blue does not expedite change, no such factor enters $\Pi_i(blue)$. The switching threshold of individual i corresponds to the lowest value of f_{green} such that the expected payoff from choosing green is larger than the expected payoff from choosing blue, i.e., we can solve $\Pi_i(green) = \Pi_i(blue)$ for f_{green}. Noting that $f_{blue} = 1 - f_{green}$, the switching threshold for individual i is given by $f_i = 0.5 - 0.5b/p(1 + \gamma_i)$.

Individuals who do not believe they can expedite change by deviating from the norm are characterized by $\gamma_i = 0$. We define the social tipping threshold f_{TT} as the fraction of individuals who have abandoned the norm such that even individuals with $\gamma_i = 0$ have an incentive to follow and

abandon the norm as well. This implies that $f_{TT} = 0.5 - 0.5\, b/p$. Put differently, once the proportion of individuals choosing green has reached f_{TT}, we expect change to be self-enforcing. Note also that for individuals with $\gamma_i > 0$, we can write $f_i = f_{TT} - 0.5\gamma_i b/p$.

Given a distribution of switching thresholds f_i and the rules describing the dynamics of change, one can simulate the proportion of individuals abandoning a norm. 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 ten periods (consistent with Fig. 3 in the main manuscript) based on the assumed process of preference change, (*ii*) for each individual i, we draw γ_i from the probability distribution $N(\mu, \sigma)$ to compute the switching thresholds given by $f_i = f_{TT} - 0.5\gamma_i b/p$, where f_{TT}, b, and p follow from the treatment parameters, and (*iii*) the process of change is simulated based on the following rule: if $q(t)$ is the proportion of individuals who have abandoned the norm at the end of period t, then in period $t + 1$, all individuals with a threshold $f_i \leq q(t)$ 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 a 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*). Other studies feature "group-wide" matching (*10-12*). In the case of group-wide matching, the expected payoff from choosing green would be given by $\Pi(green) = b_{green} - f_{blue}p$ and the expected payoff from blue would be given by $\Pi(blue) = b_{blue} - f_{green}p$. Note that the squared terms disappear compared to the pairwise matching protocol, because players always incur miscoordination costs if some other player in the group chooses the opposite color. However, the value of f_{green} for which $\Pi(green) = \Pi(blue)$, i.e., the tipping threshold, is still given by the same expression $f_{TT} = 0.5 - 0.5\, b/p$. Hence, our model can be used to analyze different matching environments though the type of matching may affect subjects' expectations about the likelihood of change, i.e., the distribution of γ_i may be affected. It is possible that other forms of matching and interaction in networks change the computation of the tipping threshold. The general point of defining the tipping threshold as the point of indifference between two choices for an individual that disregards future benefits of norm deviations would, however, still apply.

Estimation of Model Parameters and Switching Thresholds

Here, we describe the estimation technique we use to calibrate our model and generate Figure 3. Following our model, the probability that a subject deviates from the established norm is given by $P(choice_t = green) = P(f_i \leq f_{green,t-1})$. That is, a subject deviates from 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 $f_i = f_{TT} - 0.5\gamma_i b/p$, we obtain $P(choice_t = g) = P\left(f_{TT} - 0.5\gamma_i b/p \leq f_{green,t-1}\right)$ which after rearranging terms equals $P(\frac{f_{TT} - f_{green,t-1}}{0.5\ b/p} \leq \gamma_i)$. Letting $\tilde{\gamma} \equiv \frac{f_{TT} - f_{green,t-1}}{0.5\ b/p}$ and noting that $\gamma_i \sim N(\mu, \sigma)$, we obtain $P\left(\frac{\tilde{\gamma} - \mu}{\sigma} \leq z\right) = P\left(z < \frac{\mu - \tilde{\gamma}}{\sigma}\right) = \Phi(\frac{\mu}{\sigma} - \frac{1}{\sigma}\tilde{\gamma})$. Notice that this is a Probit model with independent variable $\tilde{\gamma}$, 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., $-\frac{1}{\sigma} \Big/ \frac{\mu}{\sigma} * (-1) = \mu$). 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".

The estimated distribution is $\gamma_i \sim N(1.73, 1.91)$. Thus, 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. 1A. Figure S6 below shows the distribution of switching thresholds $f_i = f_{TT} - 0.5\gamma_i b/p$ implied by the estimate for the distribution of γ_i. For *TT-43*, almost all individuals have a switching threshold greater than 0 and the average 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 (b is increased from 10 to 30 in this condition), more than 50% of the individuals are expected to be willing to instigate change (negative switching thresholds). Consistent with the data, change in *TT-30* is therefore likely to occur.

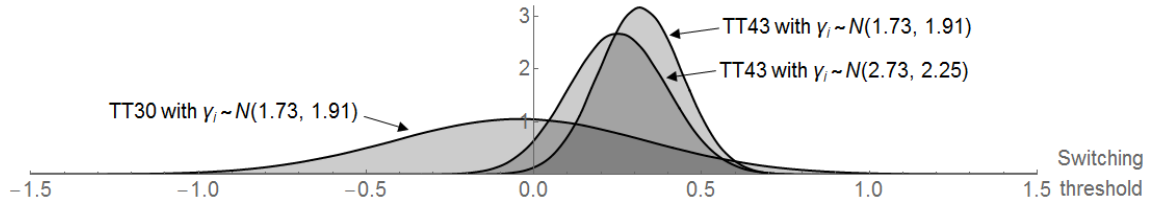

Fig. S6. Distribution of switching thresholds implied by model estimates. The estimated $\gamma_i \sim N(1.73, 1.91)$ for *TT-43* implies that almost all individuals have a switching threshold greater than 0 and the average threshold is around 35%. Further, even with $\gamma_i \sim N(2.73, 2.25)$, which 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.

4. 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. All agents have the same memory length, which, based on previous research by the same authors (*13*), is assumed to be 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 earn x experimental points if they coordinate and lose the same amount of points if they fail to coordinate. Players do not have a preference over outcomes; their only concern is to coordinate. 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. The experimental subjects are not aware of the existence/introduction of a committed minority.

The expected payoff for an individual choosing blue is given by $\Pi(blue) = f_{blue}\, x - (f_{green} + f_c)\, x$. The first term corresponds to the probability of being matched with a player who chooses blue times the benefit from coordinating. The second term captures the expected cost of failing to coordinate. The expected payoff from choosing green is computed similarly, except that individuals also take into account possible future benefits when deviating from the established convention. The expected payoff is given by $\Pi(green) = (f_{green} + f_c)x - f_{blue}\, x + \gamma_i f_c 2x$. The third term captures the future benefit. It is given by the expectation parameter γ_i times the benefit from change, which corresponds to $2x$ due to avoiding the miscoordination cost x and obtaining the coordination benefit x. Noting that $f_{blue} = 1 - f_{green} - f_c$, the value of f_{green} for which $\Pi(green) = \Pi(blue)$, i.e. 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*) actors do not have a preference for change. The impetus for change is the committed minority.

Based on the above-derived switching thresholds, we can estimate the distribution of γ_i. 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., the distribution of γ_i should be shifted 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' estimate 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 indeed corresponds to a leftward shift in the distribution compared with the estimate for our experiment (where μ = 1.73 and $\sigma = 1.91$). Moreover, Fig. S7 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 our model at predicting tipping of social conventions in a different experimental setting suggests that our model 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. S7*), 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%.

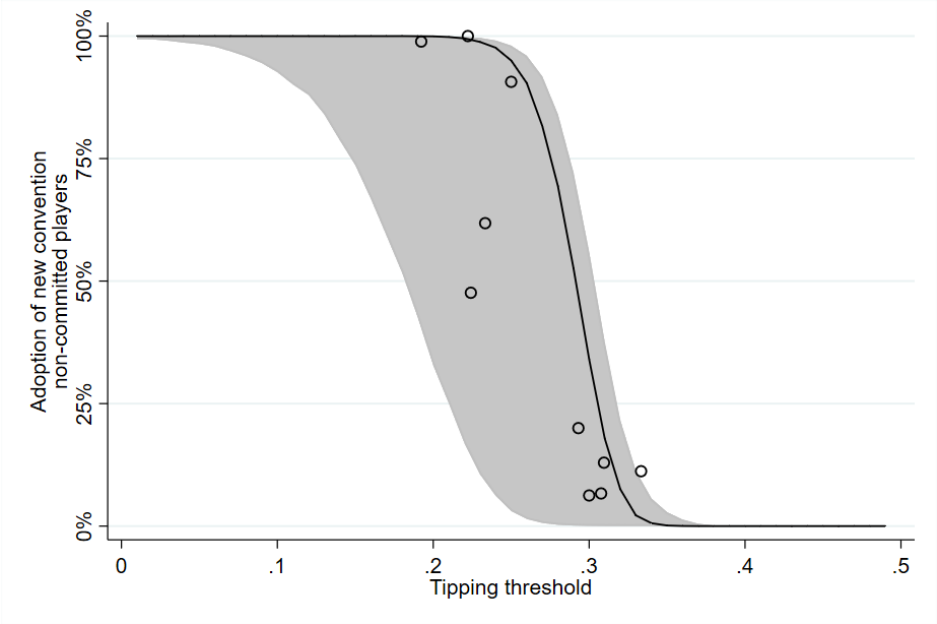

Fig. S7. Adoption of new convention for different tipping thresholds. The markers represent, for the ten experimental societies in (*8*), the percentage of choices in the last five periods that do not correspond to the initially established convention. Also shown are the theoretical predictions based on the parameters estimates μ = 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.

5. Separate Files

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

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

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

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