Predicting social tipping and norm change in controlled experiments

James Andreoni[a,1], Nikos Nikiforakis[b,c,1,2], and Simon Siegenthaler[d,1]

[a]Department of Economics, University of California San Diego, La Jolla, CA 92093; [b]Division of Social Science, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates; [c]Center for Behavioral Institutional Design, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates; and [d]Naveen Jindal School of Management, University of Texas at Dallas, Richardson, TX 75080

Edited by Ernst Fehr, University of Zurich, Zurich, Switzerland, and accepted by Editorial Board Member Paul R. Milgrom February 24, 2021 (received for review July 21, 2020)

The ability to predict when societies will replace one social norm for another can have significant implications for welfare, especially when norms are detrimental. A popular theory poses that the pressure to conform to social norms creates tipping thresholds which, once passed, propel societies toward an alternative state. Predicting when societies will reach a tipping threshold, however, has been a major challenge because of the lack of experimental data for evaluating competing models. We present evidence from a large-scale laboratory experiment designed to test the theoretical predictions of a threshold model for social tipping and norm change. In our setting, societal preferences change gradually, forcing individuals to weigh the benefit from deviating from the norm against the cost from not conforming to the behavior of others. We show that the model correctly predicts in 96% of instances when a society will succeed or fail to abandon a detrimental norm. Strikingly, we observe widespread persistence of detrimental norms even when individuals determine the cost for nonconformity themselves as they set the latter too high. Interventions that facilitate a common understanding of the benefits from change help most societies abandon detrimental norms. We also show that instigators of change tend to be more risk tolerant and to dislike conformity more. Our findings demonstrate the value of threshold models for understanding social tipping in a broad range of social settings and for designing policies to promote welfare.

social norms | conformity trap | tipping points | threshold models | laboratory experiment

Social norms are ubiquitous in human societies (1–4). By prescribing which behaviors should be rewarded and which should be punished, norms often serve the purpose of promoting welfare by discouraging harmful behaviors (e.g., smoking in public places) and encouraging beneficial ones (e.g., helping others in need). Sometimes, however, norms seem to have the opposite effect. Examples include discriminatory norms (1, 3), norms that curtail female labor force participation (5, 6), norms of personal revenge (7), and norms against same-sex marriage (8). This observation raises two critical questions: When do societies fail to abandon detrimental norms? How can policy increase the probability of abandoning them? The answers to these questions depend critically on understanding the nature of spontaneous social change, that is, change that occurs without external intervention.

A popular paradigm for modeling spontaneous change when behaviors are interdependent—as is the case when social norms exist—involves tipping points (8–16). The central idea is that social norms are backed by sanctions, which create pressure for individuals to conform to an established behavior (17–22). The pressure to conform is an essential ingredient for tipping points (13, 23). Many individuals prefer to conform if they expect others will do the same to avoid the sanctions, even if a norm change would be socially beneficial. If a critical number of individuals abandon the norm, however, the social incentives will reverse and propel rapid change toward an alternative state. From this perspective, the crucial question relating to norm change is the following: What is the critical number (or proportion) of individuals that must deviate from a norm before social incentives reverse? Put differently, when should we expect societies to reach this threshold spontaneously? The aim of this paper is to identify a theoretical model that can help answer this question.

Predicting when social tipping and norm change will occur has posed a major challenge for social scientists (15, 16): "Anyone claiming to know for sure when a particular tipping point will be reached should be treated with suspicion" (24). A striking example is the sudden disappearance of the gender gap in American higher education in the early 1970s: "The speed at which women moved from the margins to the mainstream of higher education took even knowledgeable observers by surprise" (25). While social and economic theories have identified a number of factors that can incrementally affect the likelihood of tipping, determining precisely when social tipping will occur spontaneously is difficult, as the theories either predict multiple outcomes—in which both norm abandonment and norm persistence are possible—or require specific parametric assumptions (3, 8, 10–13, 26–29). In other words, in order to predict social tipping, it is essential to have an empirically validated model. Identifying such a model seems to be vital at a time in which there are mounting calls for social change internationally, and norm change is considered by many to be essential for addressing critical

Significance

Social tipping—instances of sudden change that upend social order—is rarely anticipated and usually understood only in hindsight. The ability to predict when societies will reach a tipping point has significant implications for welfare, especially when social norms are detrimental. In a large-scale laboratory experiment, we identify a model that accurately predicts social tipping and use it to address a long-standing puzzle: Why do norms sometimes persist when they are detrimental to social welfare? We show that beneficial norm change is often hindered by a desire to avoid the costs associated with transitioning to a new norm. We find that policies that help societies develop a common understanding of the benefits from change foster the abandonment of detrimental norms.

Author contributions: J.A., N.N., and S.S. designed research; J.A., N.N. and S.S. performed research; N.N. and S.S. analyzed data; and J.A., N.N., and S.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. E.F. is a guest editor invited by the Editorial Board.

See online for related content such as Commentaries.

[1]J.A., N.N., and S.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: nikos.nikiforakis@nyu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2014893118/-/DCSupplemental.

Published April 15, 2021.

global challenges such as a loss in biodiversity and global warming (15, 23).

Empirical analysis of tipping phenomena has traditionally relied on historical (25, 29, 30) or survey data (31, 32). These studies clearly document instances of sudden social change in daily life, but the data do not permit us to identify models that can predict social tipping. Here, we present evidence from a large-scale laboratory experiment designed to test the theoretical predictions of a social tipping model. As a setting for our test, we consider one in which societal preferences change over time. As we discuss in more detail below, the cause for this change in daily life may be the arrival of new information about the alternatives, migration, or generational shifts. Irrespective of the cause, the change in societal preferences forces individuals to weigh their benefit from deviating from the social norm against the cost from not conforming to the behavior of others. We are interested to know under what conditions the change in societal preferences will lead to behavioral changes in societies and under what conditions detrimental norms will persist.

To derive testable predictions, we build on threshold models that are widely used in the theoretical literature to study norm change (8–13). A significant advantage of threshold models is that they are more tractable when analyzing dynamic systems with substantial heterogeneity of preferences (e.g., for risk or conformity) than game-theoretic models that allow for endogenous formation of expectations (8, 10). Our model allows us to derive precise predictions about when a society will abandon a detrimental norm that we can confirm or reject using laboratory experiments. At the same time, as we discuss in the concluding section, the model is general enough to be able to account for sudden changes in other contexts such as social conventions (14). The advantage of the laboratory environment is that it allows us to create the conditions necessary to test the theoretical predictions by controlling the benefit for change (e.g., how detrimental a certain norm is) as well as the cost for failing to conform to the behavior of others. In addition, laboratory experiments enable us to exogenously vary these incentives and other factors that are predicted to affect the likelihood of norm change, such as individual beliefs. Importantly, the laboratory environment allows us to replicate the same social system to ensure outcomes are not due to chance or idiosyncratic factors.

The Social Tipping Game

We design a game around three properties that are commonly discussed in the theoretical literature of norm change (8, 10). First, a social norm must exist before it can be abandoned. Second, there must be pressure to conform to the norm such that the cost of deviating is larger for instigators of change. This generates a first-mover dilemma: Even if everyone prefers change, tipping may not occur because of an incentive to wait for others to deviate first from the norm. Finally, societal preferences must evolve over time, creating an impetus for change. The laboratory environment will allow us to ensure that these conditions are satisfied and applied equally to all individuals (33).

The social tipping game is illustrated in Fig. 1. Individuals are divided into societies and interact over multiple periods. In every period, each individual is randomly matched with another one in his/her society and has to choose between two alternatives: "Blue" or "Green." Preferences over Blue and Green change over time. At the start of the game, all subjects are induced to prefer Blue: If they choose their (induced) preferred color, they earn a high reward, v_H; otherwise, they earn a low reward, v_L. By extension, they prefer coordinating on Blue to coordinating on Green. This is done for Blue to emerge as a social norm. Specifically, following refs. 1 and 8, we define a social norm as a behavioral pattern that individuals prefer to conform to on the condition that 1) most people are likely to conform to it and 2) most people believe that others ought to conform to it as well.

To satisfy condition 1, if two matched individuals fail to coordinate on the same color, they suffer a penalty. The penalty is increasing in the number of people in the society selecting the other color. Specifically, as can be seen in Fig. 1, the cost of miscoordination is $p*g_t$ for an individual choosing Blue and $p*(1-g_t)$ for an individual choosing Green, where p is a penalty parameter and g_t is the proportion of individuals choosing Green. Therefore, instigators of norm abandonment suffer a disproportionate cost (see *SI Appendix*, section 1 for details).

To check whether condition 2 is satisfied, we collected data from an incentivized questionnaire asking individuals their normative views and expectations about whether Blue or Green (or neither) is the right/most ethical/socially most appropriate choice at different points in the game. We find that condition 2 is satisfied for Blue for 98% of respondents at the start of the game (*SI Appendix*, section 2). The incentives, therefore, succeed in creating conditions such that Blue is likely to emerge as a social norm at the start of the experiment.

To generate the impetus for change, individuals' preferences shift gradually over time. To ensure we obtain precise predictions, in line with ref. 10, the process and rate at which preferences switch is public knowledge, thus ruling out pluralistic ignorance as a reason for not observing norm change (8, 10, 34). In particular, it is public knowledge that each individual has a 10% probability of experiencing a preference switch—from Blue to Green—in each period, such that after a number of periods nearly everyone prefers Green, that is, after a preference switch, individuals receive a higher payoff when coordinating on Green (Fig. 1). The Blue behavior, therefore, becomes detrimental (inefficient), in the sense that societies would benefit from change. In other words, the change in societal preferences gradually reverses the normative injunction of choosing Blue (4). This means that, when referring to "detrimental norms," we are referring to behavioral patterns for which condition 2 above has ceased to apply. In line with this, responses to our incentivized questionnaire illustrate that most individuals believe others ought to choose Green when the majority of the society's members switches preferences (*SI Appendix*, section 2). If a sufficient number of people deviate from Blue by choosing Green, then the latter can emerge as a new social norm. The emergence of Green as a norm, however, may be hindered by the disproportionate cost suffered by instigators of change and the history of adherence to the Blue norm, which affects individual expectations (1, 3).

Like with all models, different meaning can be attached to the variables in our game. The most natural interpretation of changing preferences in our setting is to think of them as modeling the gradual arrival of new information about better social alternatives. An obvious example is smoking, in which individuals over time learn about the adverse effects of cigarette consumption. A different interpretation is to think of changing preferences resulting from migration, such as new individuals arriving in a society, thus altering either directly or indirectly (through communication/imitation) the distribution of societal preferences over outcomes. Yet another interpretation is to think of them as reflecting changes due to generational shifts in preferences. Older citizens are gradually replaced with younger ones who may have greater access or openness to more recent information. Irrespective of the interpretation, however, the change of societal preferences creates the need for norm change. A similar point can be made about the interpretation of incentives in the model. The desire to conform in daily life, for example, can be due to individuals fearing sanctions, because of social image concerns, or because they have internalized the norm. Although important, this distinction will be moot in our game where our primary interest is to ensure that a pressure to conform to a norm exists, and evolving societal preferences create a need for change, such that we can test our theoretical predictions in a relevant setting.

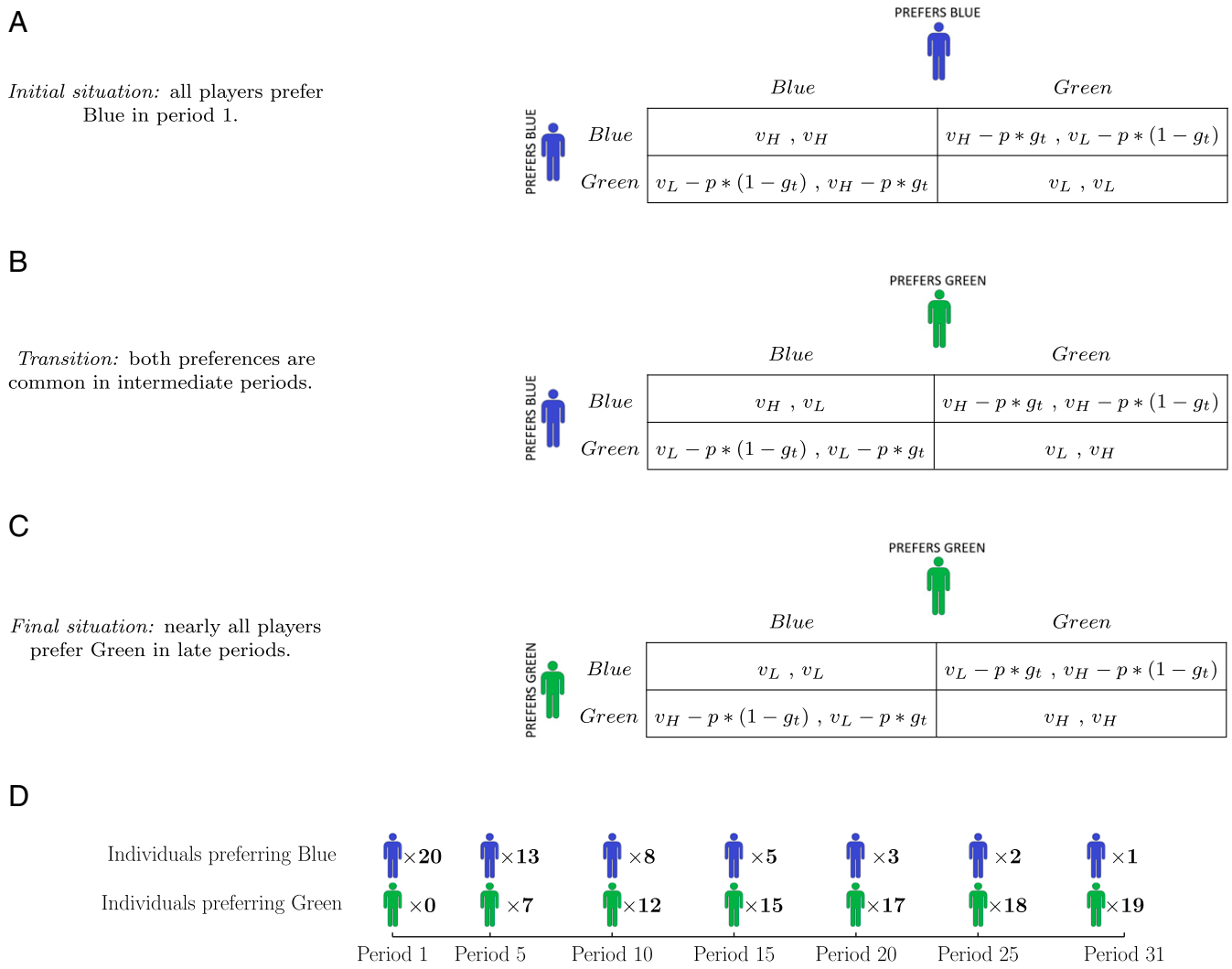

Fig. 1. Social tipping game. Individuals are divided into experimental societies of 20 individuals in all but one condition. In each of the 31 periods, individuals are randomly matched into pairs with another society member and must choose between action Blue and Green, simultaneously and without communication. If they choose their (induced) preferred color, they earn a high reward, v_H; otherwise, they earn a low reward, v_L. Individuals in a pair have an incentive to coordinate on a color. Specifically, there is cost for miscoordinating, which is given by a penalty parameter, p, multiplied by the proportion of individuals in the society choosing the other color, where g_t denotes the proportion of people choosing Green in period t. Induced preferences over Blue and Green change over time at a commonly known rate: An individual who still prefers Blue has a 10% probability of switching to preferring Green in each period. The changing preferences create three possible situations that are captured in *A*–*C*. (*A*) In period 1, it is common knowledge that everyone prefers Blue. (*B*) After the first period, there is a probability that players with conflicting preferences are paired together. (*C*) By the end of the experiment, virtually all individuals prefer Green over Blue. (*D*) This panel illustrates the process of changing preferences by presenting the number of individuals expected to prefer Blue/Green in selected periods.

Finally, we note that our game induces convergence to a norm that is initially seen as beneficial prior to becoming detrimental. This is consistent with situations in which a certain behavior was considered desirable/justified when it first emerged, a behavior which by today's standards seems undesirable (e.g., smoking, discriminatory norms, and bans on same-sex marriage). Our comparative statics would not be affected if we allowed a minority of citizens to hold opposing views either initially or later on. Of course, some norms emerge even though they are detrimental for a majority of citizens (1, 7). While our study does not help explain why such norms emerge, our findings will still be informative about the process of abandoning such norms as long as their persistence is linked to concerns for conformity and high costs for initial transgressors.

Theoretical Framework

To derive testable predictions for social tipping, we build on threshold models (8–13). Threshold models are grounded on the assumption that an individual's willingness to deviate from a norm depends on the proportion of others in the society that previously deviated from it (the individual's threshold). Individuals are assumed to have different thresholds, which can be due to different preferences over outcomes, different attitudes toward risk and conformity, or different expectations. Such heterogeneity is at the heart of the model as it influences who is willing to instigate change and who is willing to follow. The literature has emphasized the key role played by the former—"the instigators" (10), "the norm entrepreneurs" (3), "the trendsetters" (8), "the committed minority" (14), and "the great" (35)—and the need to understand what drives them.

Individual thresholds are typically taken as given (8–13) and assumed to represent a point of indifference when comparing the benefits and costs of a deviation from the status quo. In line with this, we assume that individuals will deviate from Blue as soon as they perceive the incentives for choosing Green to exceed those

for choosing Blue. As can be seen in Fig. 1, in any given period, the pecuniary payoff from choosing Green for an individual who prefers Green is $\pi(Green) = v_H - \boldsymbol{I}_{Miscoordination}\, p^*(1-g_t)$, and the pecuniary payoff from choosing Blue is $\pi(Blue) = v_L - \boldsymbol{I}_{Miscoordination}\, p^* g_t$. Recall that $v_H > v_L$. The indicator function $\boldsymbol{I}_{Miscoordination}$ equals 1 if the two individuals fail to coordinate on the same color and 0 otherwise. These pecuniary payoffs will be induced in the experiment to ensure all properties mentioned at the start of the previous section apply equally to all participants. However, they will be only part of an individual's perceived incentives for abandoning Blue in favor of Green.

The willingness to deviate from the status quo can also be affected by an individual's naturally occurring preference for change or belief in his/her ability to expedite norm change by acting against the status quo (8, 36, 37). To capture such heterogeneity, we assume that each individual i is characterized by a random variable $\gamma_i \sim N(\mu, \sigma)$. Specifically, for an individual preferring Green, the perceived utility (i.e., the pecuniary payoff plus the naturally occurring preference/expectation) for choosing Green is given by $\pi(Green) + \gamma_i v$, where $v \equiv v_H - v_L$ is the per-period gain in pecuniary payoff when change occurs. A positive γ_i means that individual i, in addition to the incentives implied by $\pi(Green)$, is motivated to choose Green by a personal preference (e.g., a dislike for conformity) and/or a belief that deviating from Blue will expedite norm change. A negative γ_i could be interpreted as a status quo bias. We assume γ_i is normally distributed as this approximates the random variation of many natural processes (8, 10); an alternative distribution is discussed in *SI Appendix*, section 3. Because γ_i is a modeling device capturing individual heterogeneity, it is not monetized in the experiment but will be estimated from the data.

To find an individual's switching threshold, we compare the perceived utilities for choosing Blue and Green. Specifically, we set equal the expectation of $\pi(Green) + \gamma_i v$ and $\pi(Blue)$ and solve for g_t. We find that the switching threshold for each individual i is given by $f_i = 0.5 - 0.5(1+\gamma_i)\, v/p$ (see *SI Appendix*, section 3 for details). This threshold corresponds to the proportion of others who must deviate from an established equilibrium before individual i is willing to do so as well. The switching threshold decreases in 1) the benefit–cost ratio of norm change, v/p, and 2) the variable γ_i measuring individual-specific preferences/expectations for change. Because $v > 0$, f_i is below 50% of the population. For large values of γ_i, the switching threshold f_i can become negative, which indicates that an individual is willing to be the first to deviate from the norm, a committed type (14). Similarly, $\gamma_i < 0$ indicates that individual i may persist in choosing Blue even when the proportion of others who have deviated would imply that $\pi(Green) > \pi(Blue)$. In *SI Appendix*, section 3, we show that we obtain virtually identical predictions for our experimental conditions if we assume the random variable γ_i follows the exponential distribution $Exp(1/\mu)$ such that $\gamma_i \in [0, \infty]$.

Given a distribution of thresholds, the dynamics of change are described by a simple 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 $f_i \le g_t$ abandon the norm as well. A society reaches a tipping threshold when the number of people who are deviating from Blue becomes large enough such that even individuals who do not have a personal preference for change and who do not believe that deviating from Blue will expedite norm change have an incentive to follow suit. Since these individuals are characterized by $\gamma_i = 0$, the tipping threshold is given by $f_{TT} = 0.5 - 0.5\, v/p$. As above, as $v > 0$, the tipping threshold is below 50% of the population and is decreasing in the benefit–cost ratio of norm change v/p. The tipping threshold provides us with a standardized measure for evaluating the prospects for change across different environments.

Fig. 2*A* shows the probability that a society spontaneously abandons a norm as a function of the tipping threshold. The probability of social tipping corresponds to the mean proportion of individuals who abandon the norm when simulating the above-described dynamics of change over 10,000 trials. Under plausible assumptions about γ_i, the probability of social tipping in our experiment is predicted to be 100% when the tipping threshold is below 35%. When the threshold exceeds 35%, the probability of norm abandonment is predicted to quickly drop to 0%. In other words, increases in the benefit–cost ratio of change are predicted to have nonlinear effects on the probability of social change (9, 12, 15). Note that norm change would be socially beneficial even when the threshold

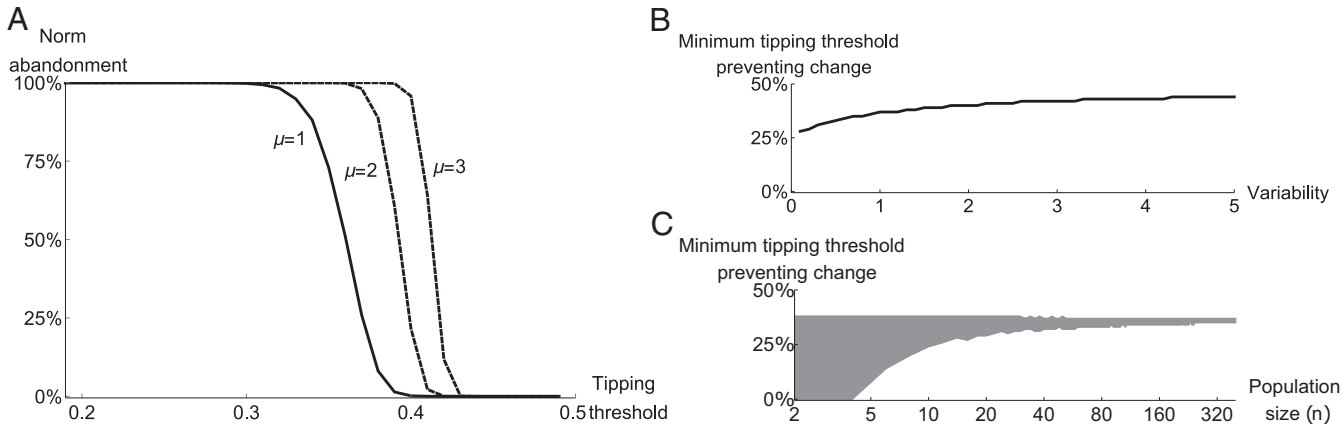

Fig. 2. Theoretically predicted norm abandonment. (*A*) Probability of norm abandonment depending on the tipping threshold. The predictions are given for $\mu = 1$ (solid line), $\mu = 2$, and $\mu = 3$ (dashed lines), assuming $\sigma = 1$. For all cases, successful change is predicted for tipping thresholds below 35%. An $\mu = 1$ seems plausible, as it implies that the average individual anticipates the existence of a tipping threshold and believes a deviation by him/her will bring society one step closer to it. Specifically, if $\gamma_i = 1$, then in addition to the myopic payoff, $\pi(Green)$, individual i associates an additional gain of $v = v_H - v_L$ for choosing Green, as i believes that society-wide coordination on the high benefit will occur one period earlier. Higher values of μ occur if the average individual has a personal preference for change or expects a deviation will accelerate change by more than one time period. (*B*) Robustness of predictions to different variability in γ_i (measured by σ) given $\mu = 1$. The minimum tipping threshold preventing change corresponds to the tipping threshold above which change is unlikely (below 50%). It lies between 30 and 40%, though the increasing trend shows that change is more likely in more heterogeneous societies. (*C*) Stability of predictions for different population sizes (n), holding constant $\mu = \sigma = 1$ (i.e., abstracting from the possibility that n could affect the distribution of γ_i). Shaded area shows the 99% CI based on 1,000 trials for each population size. The variability in the probability of change in different societies due to the stochastic nature of the model (i.e., different realized induced preferences and distributions of γ_i) is small when $n > 10$.

is above 35%, but the cost of miscoordination is such that societies are predicted to be locked into what could be described as a conformity trap. The predictions illustrate that, apart from increasing the benefits and reducing the costs, policies can aim to promote social change by inducing a change of expectations (8). Fig. 2 *A* and *B* respectively show that an increase in the mean μ and the SD σ increases the probability of social tipping. However, the increase is relatively small, indicating that the predicted drop in the probability of norm abandonment at the 35% tipping threshold is robust to small changes in expectations. Fig. 2*C* establishes that for societies consisting of 20 individuals (as is the case in our experiment) predictions exhibit little variation due to the stochastic nature of the model.

Experimental Conditions

Our laboratory sample comprises 1,020 participants divided into 54 experimental societies. To provide a thorough test of the theoretical predictions, we implemented nine experimental conditions. We describe them below (see *SI Appendix*, section 1 for details). The first four conditions explore the influence of varying the tipping threshold on the likelihood of social tipping. In particular, the baseline condition *TT-43* implements a tipping threshold of 43% for which the model predicts no social tipping (Fig. 2*A*). The parameters in this condition are $v_H = 30$, $v_L = 20$, and $p = 76$ such that $f_{TT} = 0.5 - 0.5\ v/p = 0.43$, where $v \equiv v_H - v_L$. Condition *TT-30* implements a tipping threshold of 30% due to a higher benefit of change, $v_H = 50$, and condition *TT-23* implements a tipping threshold of 23% due to a lower miscoordination penalty, $p = 19$. Since the latter thresholds are below 35%, the model predicts social tipping will occur in both conditions. Whether these predictions will be realized empirically hinges on the model's ability to capture people's innate preferences and beliefs (γ_i is not monetized in the experiment) and on the validity of the behavioral assumptions of the threshold model. Finally, in condition *TT-Endo*, subjects set the tipping threshold endogenously by choosing how much others are penalized when failing to coordinate. This is a key condition, as social norms are backed by informal sanctions (1–4, 17–22), and if individuals fail to reduce sanctions sufficiently to achieve change, it would further emphasize the need for policy intervention.

Apart from varying benefits and costs, our experiment offers an opportunity to test the efficacy of interventions that could affect beliefs/expectations about change. As explained above, the relevant expectation (captured through γ_i) is an individual's belief in his/her ability to expedite change by deviating from the norm and in his/her belief about how likely others are to follow his/her example. The second set of conditions are designed to affect such expectations, while holding the tipping threshold fixed at 43%. First, we study whether social tipping is more likely in smaller societies (*Small Society*) and when subjects receive instant information about each other's behavior (*Fast Feedback*). In *Small Society*, the size of a group is reduced to 10 individuals; hence,

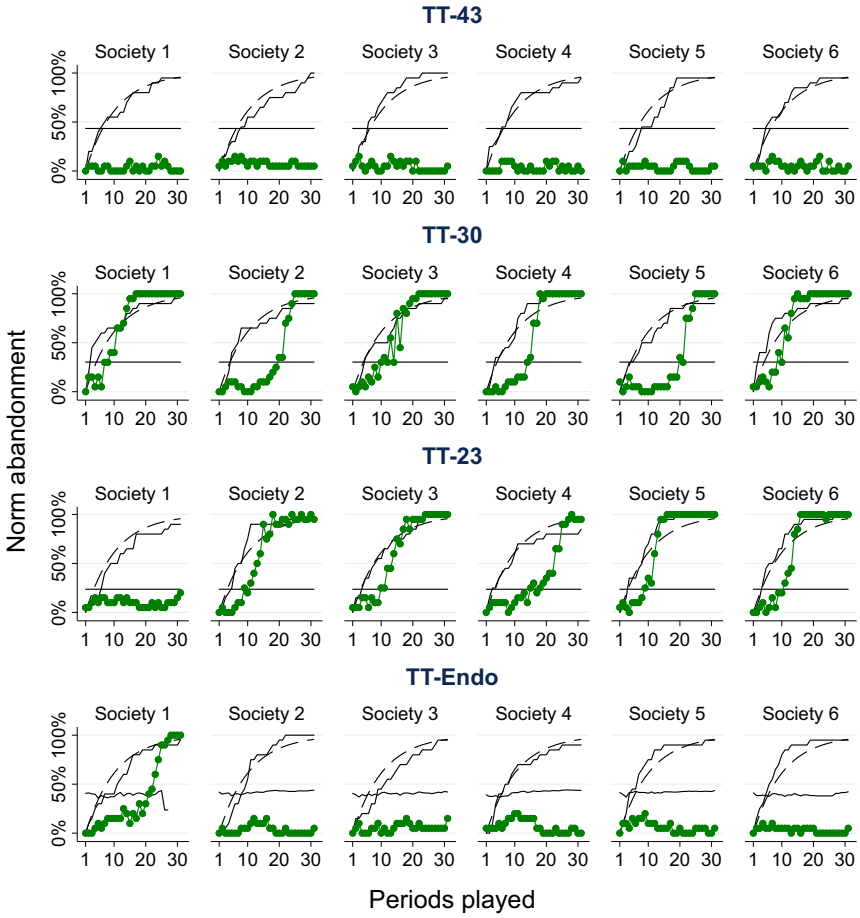

Fig. 3. Time series of norm abandonment for different tipping thresholds. Norm abandonment is shown as the line with circled markers. The tipping threshold is given by the horizontal line. The dashed concave line indicates the theoretically expected fraction of subjects preferring to abandon the norm; the solid increasing line is the corresponding realized fraction. Conditions *TT-30* and *TT-23* allow for fast and efficient change relative to *TT-43* ($P = 0.001$ and $P = 0.008$, one-sided Fisher exact test). Condition *TT-Endo* leads to an average tipping threshold of 40% and allows for change in only one out of six experimental societies ($P = 0.500$, one-sided Fisher exact test).

each individual represents a larger part of society than in the baseline condition (*TT-43*), and deviations are more impactful. In *Fast Feedback*, norm deviations are rapidly observed by others—mimicking an effect of modern-day communication. While both conditions are expected to increase an individual's belief about the prospects of change, we anticipate that their effect on the likelihood of tipping will be limited as social norms involve interdependent behaviors. The idea is that social change requires a coordinated change of expectations (8). To that end, we consider two additional conditions. In *Public Awareness*, we highlight the impetus for change by providing public information in the experimental instructions about the predominant preferences in society in any given period. Specifically, we provide information about the realized preferences in other experimental societies. In *Preference Poll*, individuals can express their preferred social alternative, Blue or Green, via a poll taking place in period 14 (i.e., when a clear majority is expected to prefer to abandon the Blue norm).

Finally, we consider an experimental condition in which we offer a reward to the four subjects in the society that chose the action that dominated at the end of the experiment for the longest time (any ties are broken randomly). The reward is designed to model social rewards commonly afforded to leaders of successful change. Accordingly, we name this condition *Incentive for Instigators*. The tipping threshold is again fixed at 43%. What makes this treatment particularly interesting is that it is difficult to predict the outcome. On the one hand, individuals have a greater incentive to instigate norm change, all else equal. On the other hand, they may be less willing to follow a leader that derives a greater benefit from change than they do. In all of the experimental conditions, we also measured participants' attitudes toward risk and conformity (see *SI Appendix*, section 1 for details).

Results

Fig. 3 depicts the time series of behavior in each society for the conditions that vary the tipping threshold. All experimental societies started by coordinating on the initially preferred behavior (Blue). Thus, Blue emerges as a social norm in all societies. In line with the predictions, when the tipping threshold was high in *TT-43*, all six societies failed to reach it, and norm change was never observed. In contrast, in the conditions with lower tipping thresholds, the fraction of individuals deviating from the established norm increased over time until the tipping threshold was reached, in which case rapid change followed: five out of six and six out of six societies achieved change in *TT-23* and *TT-30*, respectively. Strikingly, when subjects set the tipping threshold themselves by selecting the miscoordination penalty, they set it too high. On average, the tipping threshold in *TT-Endo* is 40%, almost as high as in *TT-43*, whereas subjects could have reduced it to 23% (as in *TT-23*), and five out of six societies fail to abandon the detrimental norm. In fact, we observe an increase in the miscoordination penalty over time in *TT-Endo*. This is at odds with a willingness to facilitate norm change as the penalty increases with the number of people that prefer Green in the society. Our analysis suggests that individuals increased the penalty over time to prevent the costs associated with transitioning to the new norm (*SI Appendix*, Fig. S4). We also find evidence for indirect negative reciprocity as individuals are particularly likely to raise penalties after having themselves incurred large miscoordination costs.

How well does our threshold model predict social tipping? Fig. 4 juxtaposes the theoretical predictions against the data. The model correctly predicts when a norm persists and when social tipping occurs in 96% of instances, that is, in 23 of the 24 experimental societies. We observe a sharp drop in the likelihood of change beyond a tipping threshold of 35%. This constitutes direct evidence in support of threshold models, and that varying tipping thresholds critically affects the probability of change. We also performed out-of-sample predictions to test the model. Specifically, we calibrate the model based on data from half of our experimental conditions and then show that the calibrated model continues to predict behavior accurately in the other societies (*SI Appendix*, Fig. S5).

What kind of interventions are most effective at affecting expectations for change? Fig. 5*A* shows that these are interventions which help societies coordinate expectations: In *Preference Poll*, five out of six societies achieved change; in *Public Awareness*, four out of six societies achieved change. In terms of our model, this implies an increase in γ_i of 345% in *Preference Poll* and 258% in *Public Awareness* relative to the baseline conditions (*SI Appendix*, Fig. S6). As anticipated, the other two interventions were less effective at altering expectations. Whereas change was more likely in smaller societies (*Small Society*, three out of six societies achieved change), this was not the case when societies received accelerated feedback about others' behavior (*Fast Feedback*, one out of six societies achieved change). The implied increase in γ_i is noticeably smaller in these conditions: 189% in *Small Society* and 39% in *Fast Feedback*. It is worth noting that *Small Society* yielded the lowest average earnings of all conditions, as the transition period lasts longer than in the other conditions (*SI Appendix*, Fig. S7). Also, while *Fast Feedback* led to a rapid change in one experimental society, in the other societies it reduced the number of attempts to instigate change compared to the baseline condition ($P < 0.001$, *SI Appendix*, Fig. S6). This suggests that rapid feedback can discourage instigators of change by quickly informing them that most others adhere to the existing norm.

Next, we turn our attention to individual behavior. In Fig. 5*B*, we present our analysis of the factors that influence an individual's willingness to act as an instigator of change. The negative coefficient for the tipping threshold shows that individuals are, on average, more likely to deviate from the norm if such deviations are less costly (38). In each experimental session, we elicited each subject's tolerance to risk and conformity (see *SI Appendix*, section 1 for details). Although our experiment does not allow us to establish a causal link with behavior, both measures are found to be highly correlated with one's willingness to deviate from Blue in the experiment (Fig. 5*B*). Condition *Incentive for Instigators*

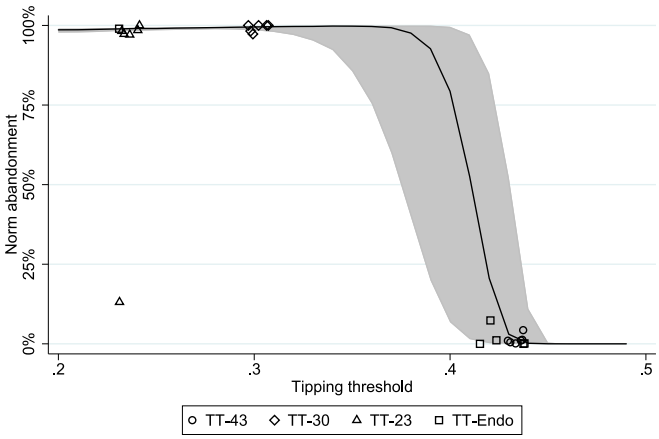

Fig. 4. Norm abandonment as a function of the tipping threshold. The tipping threshold is a critical determinant of the likelihood to observe change. Each marker represents the percentage of subjects in the last five periods that abandoned Blue in a given experimental society. Also shown is the theoretically predicted frequency of norm abandonment (solid line) and 99% CI (shaded area) from 10,000 simulated trials per tipping threshold based on the estimated parameters $\mu = 1.73$ and $\sigma = 1.91$ (Probit model with society random effects), see *SI Appendix*, section 3. The theoretical predictions correctly anticipate norm persistence or norm abandonment in 23 of the 24 societies (i.e., in 96% of instances). The model provides a similarly good fit when using a subset of the conditions to estimate μ and σ and use them to perform out-of-sample predictions (*SI Appendix*, Fig. S5).

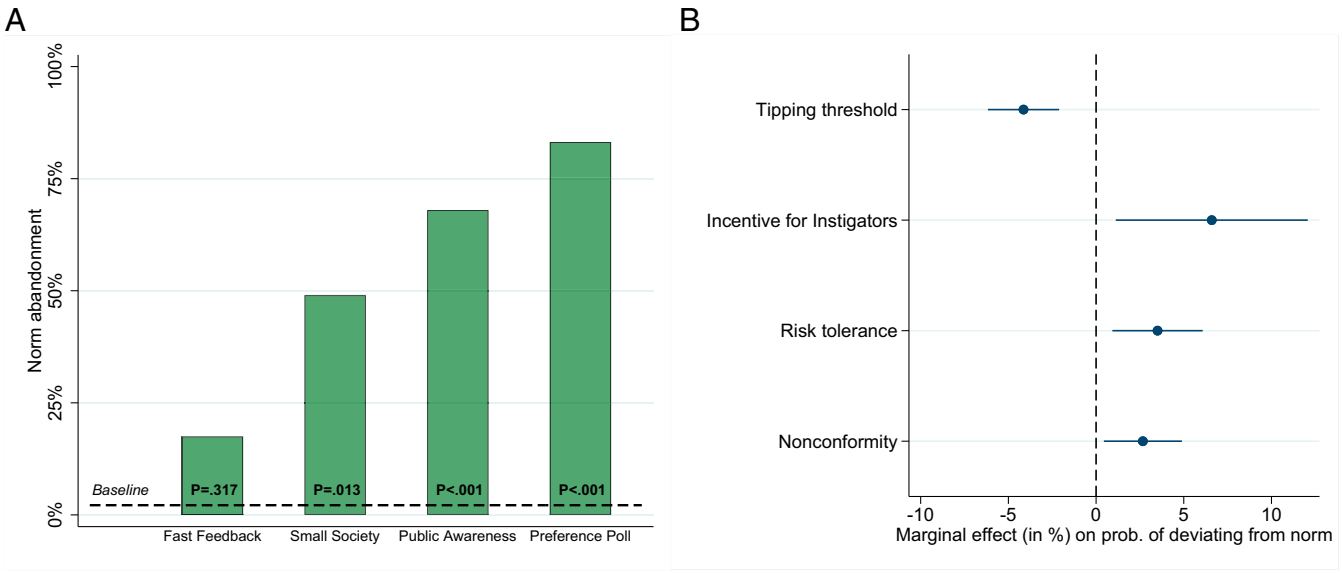

Fig. 5. Expectations and the willingness to instigate change. (*A*) Bars show the probability of norm abandonment in the last five periods in the conditions aimed to induce change via expectations. *P* values are from linear panel regressions with society-clustered SEs, in which the comparison is with *TT-43* (i.e., *TT-43* is the baseline for this comparison). In all these conditions, the tipping threshold is identical to *TT-43* (43%), showing that expectations are a crucial determinant of change. (*B*) The average marginal effects in percentage points and 99% CIs on the probability of deviating from the norm when the tipping threshold has not been reached (random effects Probit model with society-clustered SEs) are shown. Only individuals who have already experienced a preference switch are included, as individuals who prefer the status quo rarely attempt to instigate change (*SI Appendix*, Fig. S8). The higher the tipping threshold the less likely individuals are to deviate from the norm. Instigators of change tend to be more risk tolerant and more nonconformist.

generated more deviations from the norm and led to the formation of a group of instigators in all experimental societies. However, only three out of six societies eventually crossed the tipping threshold (*SI Appendix*, Fig. S6). This points to an important issue with individualized incentives to lead change: providing such incentives may motivate early instigators of change but neglects the people with a slightly lower willingness to abandon a norm. However, both are needed for social tipping. Across conditions, instigating change was a costly endeavor: The large majority of change instigators, even when change occurred, would have earned more if everyone chose Blue in all periods (*SI Appendix*, Fig. S8). This suggests that, in addition to optimistic expectations, instigators of change may have been motivated by a personal preference for social tipping, corroborating the finding that nonconformity preferences are a crucial factor for triggering change.

Finally, we explore the validity of the central behavioral assumption in threshold models, namely, that each individual has a switching threshold that characterizes his/her behavior. Using the data for each individual, we can identify bounds for their thresholds and count how many times their behavior is in accordance with them. Fig. 6*A* shows that for more than 40% of individuals, all decisions were consistent with the existence of a single switching threshold; more than two-thirds of individuals have no more than 3 out of 30 possible inconsistencies. Most of these seeming inconsistencies appear to be rationalizable. Fig. 6*B* shows that most choices that are inconsistent with a fixed threshold occur during the transition phase from the old to the new equilibrium (rounds 10 to 13), when most groups attempt to abandon the Blue norm. In line with this, 81.4% of the inconsistent choices involve an individual reverting back to Blue after having previously deviated to Green. In *SI Appendix*, Fig. S9, we show that the larger the monetary loss suffered when choosing Green (i.e., the earlier an individual deviated from Blue relative to his/her peers) the greater the likelihood that the individual reverts back to Blue. Interpreted through the lens of our model, this evidence indicates that incurring large nonconformity costs causes individuals to adjust upwards their switching thresholds, which could be attributed to them negatively updating their expectations (γ_i) about the prospects of change.

Discussion

Predicting social tipping and norm change has been a longstanding problem for social scientists. We present evidence from a large experiment in which we observe both instances of norm change and widespread persistence of detrimental social behaviors. The threshold model correctly predicts the occurrence or absence of tipping in 23 of our 24 experimental societies (i.e., in 96% of the cases). Our findings indicate that the benefit–cost ratio of norm change is a key determinant of the probability of social tipping. In addition, our experiment has provided clear evidence that societies can fail to abandon norms when they become detrimental (inefficient) without policy intervention, even under favorable conditions such as when the impetus for change is public knowledge. The evidence also indicates that effective interventions can arise from altering the benefits and costs associated with change and should aid in coordinating the expectations for change.

Although the social tipping game used in the experiment was designed to reflect incentives individuals face in the presence of norms, the insights obtained have broad implications for predicting tipping in other social settings. Threshold models have been used to study problems of collective action and also social conventions (10–12). In *SI Appendix*, section 5, we show that our model correctly predicts the occurrence or absence of tipping for all experimental societies in ref. 14, which explores the evolution of social conventions. This analysis underscores how our model and experimental game can be used for understanding social change broadly. Our study is related to ref. 14 but differs in several important dimensions, including the social domain and scope. Specifically, in addition to the coordination incentives which depend on the proportion of people choosing an action, our setting features Pareto-ranked equilibria in which one of the equilibria leads to higher returns for everyone. This creates a normative dimension fueling the desire for change (4). The lack of social tipping is most puzzling and troubling when there are Pareto-ranked

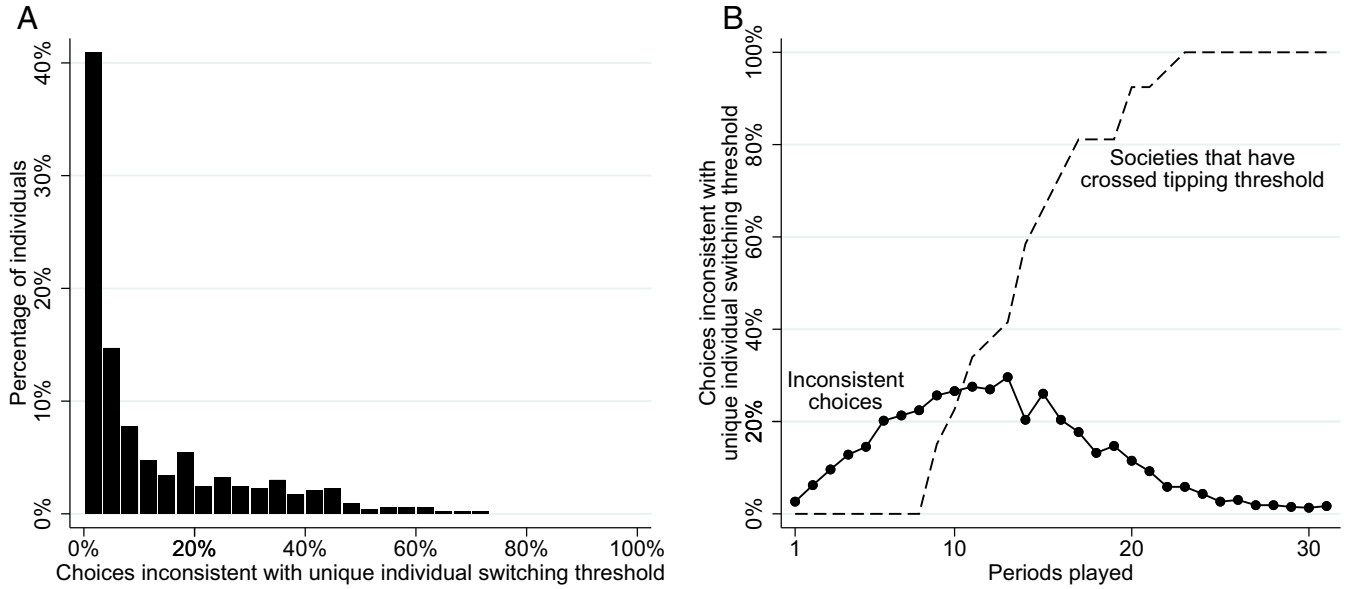

Fig. 6. Empirical validity of the key behavioral assumption of threshold models. (*A*) When an individual chooses Green for the first time, it reveals a switching threshold that is at most equal to the proportion of others who chose Green in the last observable period and at least equal to the highest proportion of others choosing Green in any prior period. A choice in any period is classified as inconsistent with a unique individual switching threshold if an individual 1) chooses Blue even though the last observed proportion of others choosing Green was greater than the upper bound of the revealed switching threshold or 2) chooses Green even though this proportion was smaller than the lower bound of the revealed switching threshold. Data only include cases in which societies eventually abandoned the Blue norm, as otherwise the switching threshold of most individuals is not observed. We find that the threshold model accurately characterizes the behavior of most individuals. (*B*) Choices that are inconsistent with the existence of a unique individual switching threshold are most common during the transition from Blue to Green, which typically gained momentum around period 10 (Fig. 3 and *SI Appendix*, Fig. S6).

equilibria. We provide clear evidence illustrating the need for and desirability of policy interventions to facilitate beneficial norm change. Unique to our study is also the fact that instigators of change emerge endogenously, allowing us to study their individual characteristics as well as the examination of different interventions for affecting expectations for social change.

Our study suggests several avenues for future research. First, it will be important to test theoretical predictions of our model on conflicting interests (9), ingroup favoritism (39, 40), the likelihood of social change when different problems compete for attention (41), and different network structures (12, 13, 42). Second, our threshold model can be used to study social change "in the wild." Specifically, it highlights what information one needs to collect to predict change. Third, our experimental setting can be used to rigorously evaluate the predictive power of statistical models for providing "early warning signals." There is an emerging field trying to identify signs of eminent tipping in ecosystems through such models (15, 16, 43, 44): "Some extremely important systems, such as the climate or ocean circulation, are singular and afford us limited opportunity to learn by studying many similar transitions" (16). The same applies to social systems (15, 24, 29). Controlled experiments are thus critical for improving our ability to detect signs that a social system is likely to tip.

Materials and Methods

The experiment was conducted at the economics laboratory of the University of California San Diego (UCSD). The experimental protocol was approved by the Institutional Review Board (IRB) at New York University (NYU) Abu Dhabi (#049-2016) and the IRB at UCSD (#150689). The main aspects of the study were explained to the subjects at the start of the experiment and consent to act as a research subject was obtained. Subjects were informed that they are allowed to opt out at all times. A total of 54 sessions were run with 1,020 subjects. Subjects were students at UCSD from various disciplines. The mean age was 20 y, and 54% of the participants were female.

Upon arriving at the laboratory, written instructions on how to make decisions in the experiment were distributed to the subjects. 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. We implemented nine experimental conditions. Each subject participated in one condition only. After the main experiment, we elicited subjects' risk and nonconformity preferences. At the end of a session, subjects were privately paid in cash. All 31 periods of the experiment were paid. The exchange rate from experimental earnings to US Dollars was $0.03 per point earned in the experiment. Payments averaged $36.10 per subject, including a show-up fee of $10. Sessions lasted less than 75 min. In *SI Appendix*, section 1, we provide the details of the experimental procedures, subjects' experience during the experiment, and the different experimental conditions.

Data Availability. Experimental data and instructions have been deposited in OpenICPSR (https://doi.org/10.3886/E134021V4) (45). All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. We gratefully acknowledge research assistance of Vincent Leah-Martin, Seung-Keun Martinez, and Alex Kellogg. We thank Andrea Baronchelli, Cristina Bicchieri, Charles Efferson, Eugen Dimant, Ernst Fehr, Karine Nyborg, Antonio Penta, Blaine Robbins, Martin Scheffer, and Roberto Weber for helpful comments. J.A. gratefully recognizes the financial support of the NSF, Grants SES-1658952 and SES-1951167. N.N. gratefully recognizes financial support by Tamkeen under the NYU Abu Dhabi Research Institute Award CG005.

1. C. Bicchieri, *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge University Press, 2006).
2. E. Fehr, U. Fischbacher, Social norms and human cooperation. *Trends Cogn. Sci.* **8**, 185–190 (2004).
3. H. P. Young, The evolution of social norms. *Annu. Rev. Econ.* **7**, 359–387 (2015).
4. E. Fehr, I. Schurtenberger, Normative foundations of human cooperation. *Nat. Hum. Behav.* **2**, 458–468 (2018).
5. M. Bertrand, E. Kamenica, J. Pan, Gender identity and relative income within households. *Q. J. Econ.* **130**, 571–614 (2015).
6. L. Bursztyn, T. Fujiwara, A. Pallais, 'Acting wife': Marriage market incentives and labor market investments. *Am. Econ. Rev.* **107**, 3288–3319 (2017).
7. J. Elster, Social norms and economic theory. *J. Econ. Perspect.* **3**, 99–117 (1989).
8. C. Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms* (Oxford University Press, 2016).

9. T. C. Schelling, *Micromotives and Macrobehavior* (WW Norton & Company, New York, 1978).
10. M. Granovetter, Threshold models of collective behavior. *Am. J. Sociol.* **83**, 1420–1443 (1978).
11. P. Oliver, G. Marwell, R. Teixeira, A theory of the critical mass: I. Interdependence, group heterogeneity, and the production of collective action. *Am. J. Sociol.* **91**, 522–556 (1985).
12. M. W. Macy, Chains of cooperation: Threshold effects in collective action. *Am. Sociol. Rev.* **56**, 730–747 (1991).
13. C. Efferson, S. Vogt, E. Fehr, The promise and the peril of using social influence to reverse harmful traditions. *Nat. Hum. Behav.* **4**, 55–68 (2020).
14. D. Centola, J. Becker, D. Brackbill, A. Baronchelli, Experimental evidence for tipping points in social convention. *Science* **360**, 1116–1119 (2018).
15. M. Scheffer *et al*., Anticipating critical transitions. *Science* **338**, 344–348 (2012). Correction in: *Science* **338**, 1029 (2012).
16. M. Scheffer *et al*., Early-warning signals for critical transitions. *Nature* **461**, 53–59 (2009).
17. E. Fehr, S. Gächter, Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
18. O. Gürerk, B. Irlenbusch, B. Rockenbach, The competitive advantage of sanctioning institutions. *Science* **312**, 108–111 (2006).
19. S. Gächter, E. Renner, M. Sefton, The long-run benefits of punishment. *Science* **322**, 1510 (2008).
20. L. Balafoutas, N. Nikiforakis, B. Rockenbach, Direct and indirect punishment among strangers in the field. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15924–15927 (2014).
21. L. Balafoutas, N. Nikiforakis, B. Rockenbach, Altruistic punishment does not increase with the severity of norm violations in the field. *Nat. Commun.* **7**, 13327 (2016).
22. L. Molleman, F. Kölle, C. Starmer, S. Gächter, People prefer coordinated punishment in cooperative interactions. *Nat. Hum. Behav.* **3**, 1145–1153 (2019).
23. K. Nyborg *et al*., Social norms as solutions. *Science* **354**, 42–43 (2016).
24. Reaching a tipping point. *Nature* **441**, 785 (2006).
25. S. Jones, Dynamic social norms and the unexpected transformation of women's higher education, 1965-1975. *Soc. Sci. Hist.* **33**, 247–291 (2009).
26. W. A. Brock, S. N. Durlauf, Discrete choice with social interactions. *Rev. Econ. Stud.* **68**, 235–260 (2001).
27. L. E. Blume, W. A. Brock, S. N. Durlauf, R. Jayaraman, Linear social interactions models. *J. Polit. Econ.* **123**, 444–496 (2015).
28. D. Acemoglu, M. O. Jackson, History, expectations, and leadership in the evolution of social norms. *Rev. Econ. Stud.* **82**, 423–456 (2014).
29. T. Kuran, The East European revolution of 1989: Is it surprising that we were surprised? *Am. Econ. Rev.* **81**, 121–125 (1991).
30. R. Amato, L. Lacasa, A. Díaz-Guilera, A. Baronchelli, The dynamics of norm change in the cultural evolution of language. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 8260–8265 (2018).
31. R. M. Kanter, Some effects of proportions on group life: Skewed sex ratios and responses to token women. *Am. J. Sociol.* **82**, 965–990 (1977).
32. J. C. Castilla-Rho, R. Rojas, M. S. Andersen, C. Holley, G. Mariethoz, Social tipping points in global groundwater management. *Nat. Hum. Behav.* **1**, 640–649 (2017).
33. V. L. Smith, Experimental economics: Induced value theory. *Am. Econ. Rev.* **66**, 274–279 (1976).
34. D. Smerdon, T. Offerman, U. Gneezy, 'Everybody's doing it': On the persistence of bad social norms. *Exp. Econ.* **23**, 392–420 (2020).
35. M. Olson, *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard University Press, 1965).
36. B. Klandermans, Mobilization and participation: Social-psychological expansions of resource mobilization theory. *Am. Sociol. Rev.* **49**, 583–600 (1984).
37. A. Bandura, *Self-Efficacy: The Exercise of Control* (Macmillan, 1997).
38. M. Mäs, H. H. Nax, A behavioral study of "noise" in coordination games. *J. Econ. Theory* **162**, 195–208 (2016).
39. C. Efferson, R. Lalive, E. Fehr, The coevolution of cultural groups and ingroup favoritism. *Science* **321**, 1844–1849 (2008).
40. D. Harris, B. Herrmann, A. Kontoleon, J. Newton, Is it a norm to favour your own group? *Exp. Econ.* **18**, 491–521 (2015).
41. M. Scheffer, F. Westley, W. Brock, Slow response of societies to new problems: Causes and costs. *Ecosystems (N. Y.)* **6**, 493–502 (2003).
42. D. Centola, A. Baronchelli, The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1989–1994 (2015).
43. V. Dakos *et al*., Slowing down as an early warning signal for abrupt climate change. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 14308–14312 (2008).
44. J. Jiang *et al*., Predicting tipping points in mutualistic networks through dimension reduction. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E639–E647 (2018).
45. J. Andreoni, N. Nikiforakis, S. Siegenthaler, Data for "Predicting social tipping and norm change in controlled experiments". *MI: Inter-university Consortium for Political and Social Research [distributor]*, https://doi.org/10.3886/E134021V4 (2021) Published ahead of print.