NBER WORKING PAPER SERIES

PREDICTING SOCIAL TIPPING AND NORM CHANGE IN CONTROLLED EXPERIMENTS

James Andreoni
Nikos Nikiforakis
Simon Siegenthaler

Working Paper 27310
http://www.nber.org/papers/w27310

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2020

The authors gratefully acknowledge research assistance of Vincent Leah-Martin, Seung-Keun Martinez, and Alex Kellogg. The authors thank Andrea Baronchelli, Cristina Bicchieri, Eugen Dimant, Ernst Fehr, Karine Nyborg, Antonio Penta, Blaine Robbins and Martin Scheffer for helpful comments. Andreoni gratefully recognizes the financial support of the National Science Foundation, Grant 1658952. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Predicting Social Tipping and Norm Change in Controlled Experiments
James Andreoni, Nikos Nikiforakis, and Simon Siegenthaler
NBER Working Paper No. 27310
June 2020
JEL No. D03,D91,H00

ABSTRACT

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 due to the lack of experimental data for evaluating competing models. We present evidence from a large-scale lab experiment designed to test the theoretical predictions of a threshold model for social tipping. 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 predicts accurately social tipping and norm change in 96% of experimental societies. Strikingly, when individuals determine the cost for non-conformity themselves, they set it too high, causing the persistence of 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 designing policies to promote welfare.

James Andreoni
Department of Economics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
and NBER
andreoni@ucsd.edu

Nikos Nikiforakis
New York University Abu Dhabi
Social Science Division
P.O. Box 129 188
Abu Dhabi, United Arab Emirates
nikos.nikiforakis@nyu.edu

Simon Siegenthaler
Jindal School of Management
University of Texas at Dallas
800 W. Campbell Road
Richardson, TX 75080
s.siegenthaler@gmail.com

An appendix is available at http://www.nber.org/data-appendix/w27310

Introduction

Norms form the basis for all social interactions in human societies (*1-4*). By prescribing what behaviors should be rewarded and what should be punished, social 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 who need assistance). 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 such norms? The answers to these questions depend critically on understanding the nature of spontaneous social change, i.e., 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 (*17-21*). The pressure to conform is an essential ingredient for tipping points (*13, 22*). 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 abandons the norm, however, the social incentives will reverse and propel rapid change towards 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 thus improving our understanding of the process of norm change.

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" (*23*). 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" (*24*). 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 – where both norm abandonment and norm persistence are possible – or require specific parametric assumptions that require empirical validation (*3, 8, 10-13, 25-28*). In other words, the difficulty of predicting social tipping stems from the lack of an *empirically-validated* model for understanding change. Identifying such a model seems to be vital at a time in which many consider norm change as being essential for addressing critical global challenges such as global warming and loss in biodiversity (*15, 22*).

Empirical analysis of tipping phenomena has traditionally relied on historical (*24, 28, 29*) or survey data (*30, 31*). 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 lab 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 it comes to analyzing dynamic systems with substantial heterogeneity of preferences, e.g., for risk or conformity, than game-theoretic models (*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 lab experiments. 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 factors influencing individual beliefs without affecting incentives. Importantly, the laboratory environment also allows us to replicate the same social system to ensure outcomes are not due to chance or idiosyncratic factors.

Experimental Setting

We design a novel experimental setting around four 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, pressure must exist to conform to the norm, so deviating from a norm must be costly (*17-21*). Third, the cost of deviating must be larger for instigators of change, generating a first-mover dilemma: even if everyone prefers change, tipping may not occur due to 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 lab environment allows us to ensure that these conditions are common knowledge and apply to all individuals (*32*).

Our laboratory sample comprises of 1,020 participants divided into 54 experimental societies and 9 experimental conditions. Each society consists of 20 individuals (except in one condition) that interact over multiple periods. In every period, each individual is randomly matched with another and has to choose between two alternatives: "blue" or "green". At the start of the experiment, all subjects are induced to prefer blue, as they receive a higher monetary reward for choosing blue than for choosing green. This is done for blue to emerge as a social norm. Specifically, following (*1, 8*), we define a social norm as a behavioral pattern that individuals prefer to conform to on the condition that most people (i) are likely to conform to it, and (ii) believe that others ought to conform to it as well. Blue satisfies condition (ii) at the start of the experiment as all individuals prefer blue to green (*4*). To satisfy condition (i) and to model the pressure to conform, 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. Therefore, instigators of norm abandonment suffer a disproportionate cost (see *SI Appendix, section 1* for details).

To generate the impetus for change, individuals' preferences shift gradually over time. In particular, in each period, each individual has a probability of experiencing a preference switch – from blue to green – such that after a number of periods everyone prefers green. That is, after a preference switch individuals receive a higher monetary reward for choosing green. The blue norm, 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*). If a sufficient number of people deviate from blue by choosing green, then the latter can emerge as a new social norm. To ensure we obtain precise predictions, in line with (*10*), the process and rate at which preferences switch is public knowledge, thus ruling out pluralistic ignorance as a reason for detrimental norm persistence (*8, 10, 33*). Despite this, we expect the emergence of green as a social norm will be difficult given the pressure to conform to the old norm, the disproportionate cost suffered by instigators of change, and the history of adherence to the old norm which affects individual expectations (*1, 3*).

Like with all models, different meaning can be attached to the variables in our setting. 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, where 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 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, 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 is moot in our setting where our primary interest is to ensure that the four properties listed at the start of this section are satisfied such that we can test our theoretical predictions in a relevant setting.

Finally, we note that our setting induces convergence to a norm that is initially seen as beneficial prior to becoming perceived as 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 seem undesirable e.g., smoking, discriminatory norms, 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 change in such cases as long as norm 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 based on the assumption that an individual's willingness to deviate from a norm depends on the number (or proportion) of others in the society that previously deviated from it (the individual's *threshold*). Individuals are assumed to have different thresholds due to differences in personality traits, e.g., attitudes toward risk and conformity. This heterogeneity is at the heart of the model as it influences who is willing to instigate change and who is willing to follow. Indeed, 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*), "the great" (*34*) – and the need to understand what drives them (*8*).

We use a rational-choice framework to derive individual thresholds. Specifically, we assume that individuals will deviate from the blue norm as soon as their incentives for choosing green exceed those for choosing blue. Incentives depend on the net benefit, b, from choosing their preferred color and on the (maximum) penalty, p, for failing to conform to the norm. Incentives also depend on expectations about how one's deviation from the norm affects the likelihood of successful change. The latter is sometimes referred to in the literature as "self-efficacy" (*8, 35, 36*). In particular, each individual i is characterized by a variable γ_i which captures their belief about the number of rounds by which their deviation will expedite change, compared to a situation in which they continue to abide to the norm. In other words, γ_i captures the expected benefit from instigating change. We follow previous authors in assuming $\gamma_i \sim N(\mu, \sigma)$ as this approximates the random variation of many natural processes (*8, 10*). Intuitively, we expect $\gamma_i > 0$ for most individuals, which implies that $\mu > 0$.

By comparing the incentives for choosing blue and green, we can derive for each individual i a switching threshold f_i. This threshold corresponds to the proportion of others who must deviate from an established norm before individual i is willing to do so as well. In *SI Appendix, section 3* we provide mathematical expressions for the expected payoffs and use them to derive the switching thresholds. We show that an individual's switching threshold is given by $f_i = 0.5 - 0.5(1 + \gamma_i)\, b/p$. Thus, the switching threshold decreases in (*i*) the benefit-cost ratio of norm abandonment, b/p, and *(ii)* self-efficacy, γ_i. In the context of detrimental norms we have $b > 0$, implying that f_i is below 50% of the population. For large values of γ_i, the switching threshold can become negative, which indicates that an individual is willing to be the first to deviate from the norm.

Given a distribution of switching thresholds, we can take advantage of the elegance of threshold models. The dynamics of change can be described by a simple rule: if $q(t)$ is the proportion of individuals who are believed to have abandoned the norm at the end of period t, then in period $t + 1$, all individuals with a threshold $f_i \leq q(t)$ abandon the norm as well. A *society* reaches a *tipping threshold* when the number of people who are deviating from the norm becomes large enough such that even individuals who do not believe they can expedite change have an incentive to follow suit. Since these individuals are characterized by $\gamma_i = 0$, the tipping threshold is given by $f_{TT} = 0.5 - 0.5\, b/p$. As above, if $b > 0$, the tipping threshold is below 50% of the population and is decreasing in the benefit-cost ratio of norm change b/p. The tipping threshold provides us with a standardized measure for evaluating the prospects for change across different environments.

Figure 1A shows that, 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 non-linear effects on the probability of social change (*9, 12, 15*). Note that norm change would be socially beneficial even when the threshold is below 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 collective change of expectations (*8*). Figures 1A and 1B show, respectively, that an increase in the mean μ and the standard deviation σ of the distribution of γ_i increases the probability of social tipping. However, the increase is relatively small, suggesting that the predicted drop in the probability of norm abandonment at the 35% tipping threshold is robust to small changes in expectations. Figure 1B also establishes the robustness of the predictions for different population sizes.

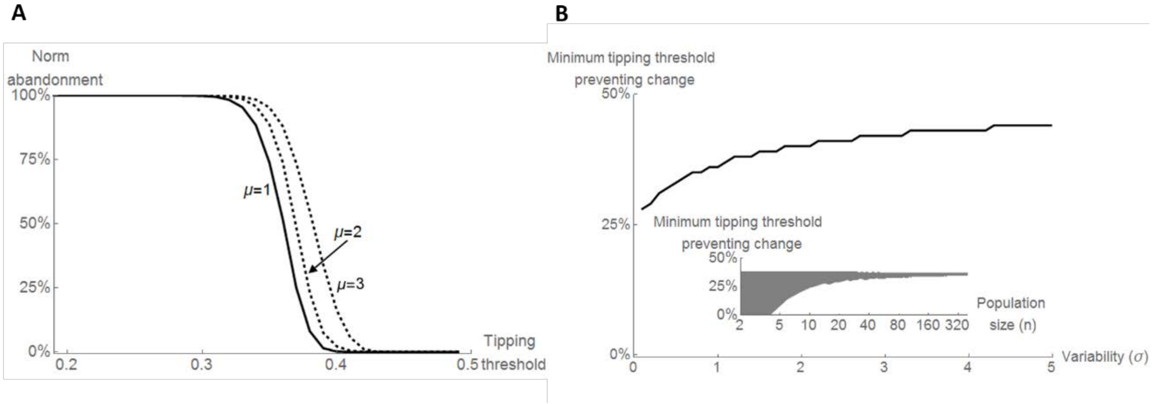

Fig. 1. Theoretically predicted norm abandonment. A) Probability of norm abandonment for different tipping thresholds. The predictions are given for $\mu = 1$ (solid line), $\mu = 2$ and $\mu = 3$ (dashed lines) assuming $\sigma = 1$. For all cases, successful change is predicted for tipping thresholds below 35%. The case $\mu = 1$ is intuitive as most individuals have neutral beliefs in the sense that they expect their deviation from the norm does not shift others' switching thresholds. **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 the probability of change is less than 50%. It generally lies between 30% and 40%, though the increasing trend shows that change is more likely in more heterogeneous societies. **Inset)** Robustness of predictions to different population sizes. Shaded area shows the 99% confidence interval based on 1,000 trials for each population size. Variations in the probability of change due to the stochastic nature of the model are small when $n \geq 10$.

Experimental Conditions

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. Condition *TT-30* implements a tipping threshold of 30% due to a higher benefit of change, b, and condition *TT-23* implements a tipping threshold of 23% due to a lower miscoordination penalty, p. Since the latter thresholds are below 35%, the model predicts social tipping will occur in both conditions. In contrast, in *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, 17-21*), and if individuals fail to reduce sanctions sufficiently to achieve change, it would further emphasize the need for policy intervention.

Apart from varying benefit and costs, our experiment offers an opportunity to test the efficacy of interventions that could affect *expectations* about change (γ_i). The second set of conditions does this 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* 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 communications. While both conditions are expected to increase self-efficacy, we anticipate that their effect on the likelihood of tipping will be limited as social norms involve interdependent behaviors. Hence, 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 "color" that dominated at the end of the experiment (i.e., blue or green) 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

Figure 2 depicts 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: 5/6 and 6/6 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, at 40%, and fail to achieve change in 5/6 societies in *TT-Endo*. 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. Our analysis suggests that individuals *increased* the penalty over time to prevent the costs associated with transitioning to the new norm (*SI Appendix, Fig. S1*). We also find evidence for indirect negative reciprocity as individuals are particularly likely to raise penalties after having themselves incurred large miscoordination costs.

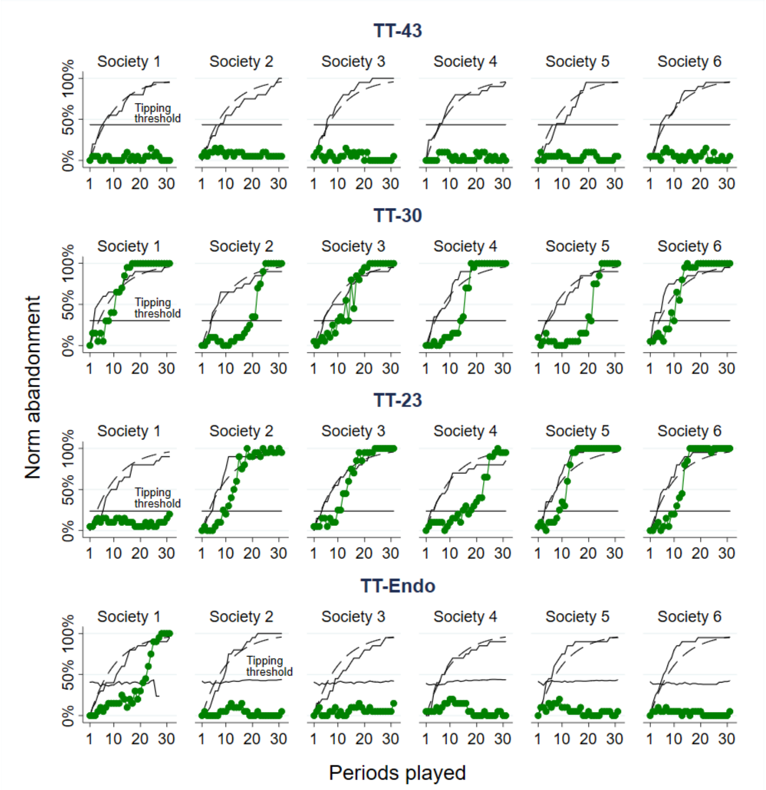

Fig. 2. Time series of norm abandonment for different tipping thresholds. Norm abandonment is shown as the line with circled markers. The tipping threshold is given by the horizontal line. The dashed concave line indicates the theoretically expected fraction of subjects preferring to abandon the norm; the solid increasing line the corresponding realized fraction. Conditions *TT-30* and *TT-23* allow for fast and efficient change relative to *TT-43*, (P=.001 and P=.008, one-sided Fisher exact test). Condition *TT-Endo* leads to an average tipping threshold of 40% and allows for change in only 1/6 experimental societies (P=0.500, one-sided Fisher exact test).

How well does our threshold model predict social tipping? Figure 3 juxtaposes the theoretical predictions against the data. The model correctly predicts when social tipping will occur 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 (estimation samples) and then show that the calibrated model continues to predict behavior accurately in the other societies (*SI Appendix, Fig. S2*).

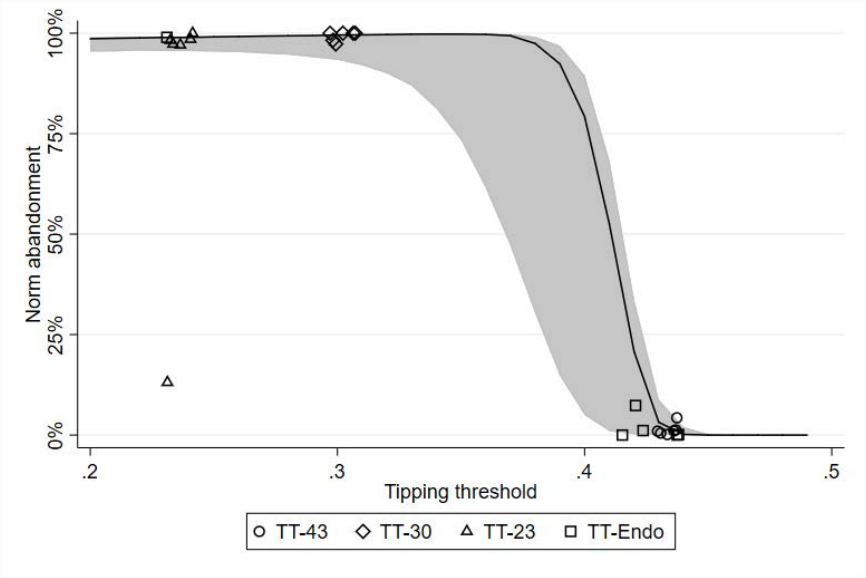

Fig. 3. Norm abandonment as a function of the tipping threshold. The tipping threshold is a critical determinant of the likelihood to observe change. Each marker represents the percentage of subjects in the last five periods that abandoned the "blue norm" for a given experimental society. Also shown is the theoretically predicted frequency of norm abandonment (solid line) and 99% confidence interval (shaded area) from 10,000 simulated trials per tipping threshold based on the estimated parameters μ = 1.73 and $\sigma = 1.91$ (Probit model with society random effects, see *SI Appendix, section 3*). The theoretical predictions correctly anticipate 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, section 4*).

What kind of interventions are most effective at affecting individuals' expectations for change? Figure 4A shows that these are interventions which help societies coordinate expectations: in *Preference Poll*, 5/6 societies achieved change; in *Public Awareness*, 4/6 societies achieved change. In terms of our model, this implies an increase in γ_i (the variable measuring the expected benefits from instigating change) of 345% in *Preference Poll* and 258% in *Public Awareness* relative to the baseline conditions (*SI Appendix, Fig. S3*). As anticipated, the other two interventions were less effective at altering expectations. Whereas change was more likely in smaller societies (*Small Society*, 3/6 societies achieved change), this was not the case when societies received accelerated feedback about others' behavior (*Fast Feedback*, 1/6 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, Figs. S3, S4*). 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<.001, *SI Appendix, Fig. S3*). This suggests that rapid feedback can discourage instigators of change by quickly informing them that most others adhere to the existing norm.

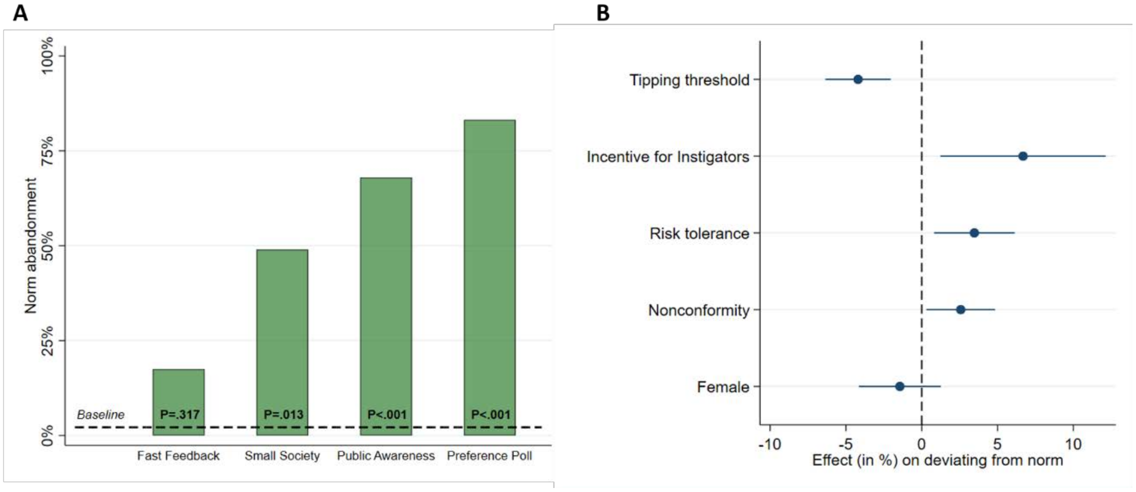

Fig. 4. Expectations and the willingness to instigate change. A) Bars show the probability of norm abandonment in the last five periods in the conditions aimed to induce change via expectations. P-values are from linear panel regressions with society-clustered standard errors, where the comparison is with *TT-43* (see supplementary text). In all these conditions the tipping threshold is identical to *TT-43* (43%), showing that expectations are a crucial determinant of change. **B)** Shown are the average marginal effects and 99% confidence intervals for the probability of deviating from the norm when the tipping threshold has not been reached (random effect Probit model with society-clustered standard errors). 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. S5*). 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 non-conformist.

Finally, in Fig. 4B, we present our analysis on the factors that influence individuals' willingness to act as instigators of change. In each experimental session, we elicited subjects' risk tolerance and preference for nonconformity (*SI Appendix, section 1*). Both measures are found to be highly correlated with one’s willingness to deviate from the blue norm in the experiment. On the other hand, there is no significant difference between men and women in their willingness to deviate from the blue norm. Condition *Incentive for Instigators* generated more deviations from the norm and led to the formation of a group of instigators in all experimental societies. However, only 3/6 societies eventually crossed the tipping threshold (*SI Appendix, Fig. S3*). 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. Viewed over all 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. S5*). This suggests that instigators of change were motivated by a personal preference for social tipping, corroborating the finding that nonconformity preferences are a crucial factor for triggering change.

Discussion

Predicting social tipping has been a long-standing problem for social scientists due to the lack of empirically-validated models. Our experimental data shows both instances of norm persistence and norm tipping. The threshold model correctly predicts the occurrence 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 detrimental norms without policy intervention, even under favorable conditions such as when the impetus for change is public knowledge. The evidence also indicates that effective interventions, apart from altering the benefits and cost associated with change, should aim to coordinate social expectations for change.

Although our analysis is conducted in the context of social 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 4* we show that our model correctly predicts the occurrence of tipping for all experimental societies in (*14*) who explore the evolution of social conventions. This analysis underscores how our model and experimental setting can be used for understanding social change broadly. Our study is related to (*14*), but differs in several important dimensions, including the social domain and scope. Regarding the social domain, we explore tipping in social norms. In addition to the coordination incentives, in our setting there is a clear normative dimension as one of the actions leads to higher returns for everyone (*4*). It is in these instances in which the lack of social tipping is most puzzling and troubling. Regarding the scope, we provide the first 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 interesting avenues for future research. First, it will be interesting to test experimentally the accuracy of novel theoretical predictions of our model about the likely impact on social change of conflicting interests (*9*), ingroup favoritism (*37*), the complexity of change when different problems compete for attention (*38*), and different network structures (*12, 13, 39*). 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. This seems like an important continuation of our work. Third, our new 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, 40, 41*): "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, 23, 28*). 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 IRB at NYU Abu Dhabi (#049-2016) and the IRB at UCSD (#150689). A total of 54 sessions was run with 1020 subjects. Each subject participated in one session only. Subjects were students at UCSD from various disciplines. The mean age was 20 years 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 sheet. Subjects interacted via computer terminals. We implemented nine experimental conditions, whose main features are described above. After the main experiment, we continued by eliciting subjects' risk and nonconformity preferences. At the end of a session, subjects were privately paid in cash. Payments averaged $36.10 per subject, including a show up fee of $10. Sessions lasted less than 75 minutes. In *SI Appendix, section 1* we provide the details of the experimental procedures, subjects' experience during the experiment and the different experimental conditions.

Acknowledgments

The authors gratefully acknowledge research assistance of Vincent Leah-Martin, Seung-Keun Martinez, and Alex Kellogg. The authors thank Andrea Baronchelli, Cristina Bicchieri, Eugen

Dimant, Ernst Fehr, Karine Nyborg, Antonio Penta, Blaine Robbins and Martin Scheffer for helpful comments. Andreoni gratefully recognizes the financial support of the National Science Foundation, Grant 1658952.

References

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 in Cognitive Sciences* 8, 185-190 (2004).

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

4. E. Fehr, I. Schurtenberger, Normative foundations of human cooperation. *Nature Human Behaviour* 2, 458-468 (2018).

5. M. Bertrand, E. Kamenica, J. Pan, Gender identity and relative income within households. *The Quarterly Journal of Economics* 130, 571-614 (2015).

6. L. Bursztyn, T. Fujiwara, A. Pallais, Acting wife: marriage market incentives and labor market investments. *American Economic Review* 107, 3288-3319 (2017).

7. J. Elster, Social norms and economic theory. *Journal of Economic Perspectives* 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. Schelling, *Micromotives and macrobehavior* (WW Norton & Company, New York, 1978).

10. M. Granovetter, Threshold models of collective behavior. *American Journal of Sociology* 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. *American Journal of Sociology* 91, 522-556 (1985).

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

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

14. D. Centola *et al.*, 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).

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. Ö. 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. *Proceedings of the National Academy of Sciences* 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. *Nature Communications* 7, 1-6 (2016).

22. K. Nyborg *et al*., Social norms as solutions. *Science* 354, 42-43 (2016).

23. Nature Editorial, Reaching a tipping point. *Nature* 441, 785 (2006).

24. S. Jones, Dynamic social norms and the unexpected transformation of women's higher education, 1965-1975. *Social Science History* 33, 247-291 (2009).

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

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

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

28. T. Kuran, The East European revolution of 1989: is it surprising that we were surprised? *American Economic Review* 81, 121-125 (1991).

29. R. Amato, L. Lacasa, A. Diaz-Guileara, A. Baronchelli, The dynamics of norm change in the cultural evolution of language. *Proceedings of the National Academy of Sciences* 115, 8260-8265 (2018).

30. R. Kanter, Some effects of proportions on group life: skewed sex ratios and responses to token women. *American Journal of Sociology* 82, 965-990 (1977).

31. J. Castilla-Rho *et al.*, Social tipping points in global groundwater management. *Nature Human Behaviour* 1, 640-649 (2017).

32. V. Smith, Experimental economics: Induced value theory. *The American Economic Review* 66, 274-279 (1976).

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

34. M. Olson, *The logic of collective action: public goods and the theory of groups* (Harvard University Press, 1965).

35. B. Klandermans, Mobilization and participation: social-psychological expansions of resource mobilization theory. *American Sociological Review* 49, 583-600 (1984).

36. A. Bandura, *Self-efficacy: the exercise of control* (Macmillan, 1997).

37. C. Efferson, R. Lalive, E. Fehr, The coevolution of cultural groups and ingroup favoritism. *Science* 321, 1844-1849 (2008).

38. M. Scheffer, F. Westley, W. Brock, Slow response of societies to new problems: causes and costs. *Ecosystems* 6, 493-502 (2003).

39. 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).

40. V. Dakos *et al.*, Slowing down as an early warning signal for abrupt climate change. *Proceedings of the National Academy of Sciences* 105, 14308-14312 (2008).

41. J. Jiang *et al.*, Predicting tipping points in mutualistic networks through dimension reduction. *Proceedings of the National Academy of Sciences* 115, E639-E647 (2018).