American Economic Review 2020, 110(11): 3522–3548
https://doi.org/10.1257/aer.20171175

From Extreme to Mainstream: The Erosion of Social Norms[†]

By Leonardo Bursztyn, Georgy Egorov, and Stefano Fiorin*

Social norms, usually persistent, can change quickly when new public information arrives, such as a surprising election outcome. People may become more inclined to express views or take actions previously perceived as stigmatized and may judge others less negatively for doing so. We examine this possibility using two experiments. We first show via revealed preference experiments that Donald Trump's rise in popularity and eventual victory increased individuals' willingness to publicly express xenophobic views. We then show that individuals are sanctioned less negatively if they publicly expressed a xenophobic view in an environment where that view is more popular. (*JEL* D72, D85, Z13)

Social norms, the set of "social sanctions or rewards" that incentivize a certain behavior (Bénabou and Tirole 2011), are an important element of any society: some behaviors and opinions are socially desirable, while others are stigmatized. There is growing evidence that individuals care to a large extent about how they are perceived by others and that such concerns might affect important decisions in a variety of settings, from charitable donations (Andreoni and Bernheim 2009; DellaVigna, List, and Malmendier 2012; Andreoni, Rao, and Trachtman 2017) to schooling choices (Bursztyn and Jensen 2015) to political behavior (Gerber, Green, and Larimer 2008; DellaVigna et al. 2017; Enikolopov et al. 2017; Perez-Truglia and Cruces 2017). Moreover, these social image concerns matter both in interactions with other people from the same social group (Bursztyn and Jensen 2015) and in interactions with strangers, such as surveyors and solicitors (DellaVigna, List, and Malmendier 2012; DellaVigna et al. 2017).

*Bursztyn: University of Chicago and NBER (email: bursztyn@uchicago.edu); Egorov: Kellogg School of Management, Northwestern University and NBER (email: g-egorov@kellogg.northwestern.edu); Fiorin: Bocconi University and IGIER (email: stefano.fiorin@unibocconi.it). Roland Bénabou was the coeditor for this article. We thank four anonymous referees, Daron Acemoglu, Abhijit Banerjee, Davide Cantoni, Esther Duflo, Benjamin Enke, Raymond Fisman, Tarek Hassan, John List, Emir Kamenica, Ricardo Perez-Truglia, Frank Schilbach, Andrei Shleifer, Hans-Joachim Voth, Noam Yuchtman, and numerous seminar participants for helpful comments and suggestions. Excellent research assistance was provided by Raymond Han, Alena Kang-Landsberg, Andrew Kao, Jacob Miller, Aakaash Rao, Giacomo Stazi, Parker Whitfill, and Rebecca Wu. We are grateful to the UCLA Behavioral Lab for financial support. This study received approval from the UCLA and U Chicago Institutional Review Boards. The experiments can be found in the AEA RCT Registry (AEARCTR-0001752 and AEARCTR-0002028).

[†]Go to https://doi.org/10.1257/aer.20171175 to visit the article page for additional materials and author disclosure statements.

A recent literature has documented the persistence of cultural traits and norms over long periods of time (Voigtländer and Voth 2012; Fernández 2007; Giuliano 2007; Algan and Cahuc 2010; Alesina, Giuliano, and Nunn 2013). However, little is known about what factors might lead long-standing social norms to *change*, or even more so, to change *quickly*. In this paper, we argue that aggregators of private opinions in a society, such as elections, might lead to updates in individuals' perceptions of what people around them think, and thus induce fast changes in the social acceptability of holding and expressing certain opinions. This, in turn, might increase the likelihood that these opinions are publicly expressed and reduce the extent to which these expressions are negatively judged and sanctioned by others.[1]

Consider the support for the communist regime in the Soviet Union in the late 1980s. Kuran (1991) argues that many individuals opposed the regime but believed that others supported it. In that environment, a referendum on the regime would have quickly updated people's opinions about the views of others. Incorrect beliefs about the opinions of others are not restricted to totalitarian regimes, where expressing personal views is often risky. In fact, as we argue below, if most individuals assume that a specific opinion is stigmatized, the stigma may be sustained in equilibrium.[2]

In this paper, we examine how social norms can be eroded quickly when new public information arrives naturally, such as an election outcome.[3] We use experiments to test the idea that Donald Trump's rise in popularity and eventual victory in the 2016 US Presidential election causally increased individuals' willingness to publicly express anti-immigration (xenophobic) views, as well as the social acceptability of such expression.[4] We examine both sides of a social interaction: agents engaging in observable behavior and observers, judging and potentially sanctioning/rewarding the behavior of agents. Our first experiment indicates that Donald Trump's rise in popularity and eventual victory in the 2016 US Presidential election increased individuals' willingness to engage in public xenophobic behavior. We also examine how this process led to changes on the *observer*'s side as well. We document that even

[1] In the view of social norms we adopt, they guide public or potentially public, but not private, actions. We do not take a broader view on social norms that also includes self-image concerns that can shape even one's private behavior by rewarding adherence and punishing deviance. Notice that since our paper explains how new public information can change social norms in the narrower (and our preferred) sense, it also suggests that social norms in the broader sense change: specifically, the part of social norms responsible for rewards or punishments by *others*. Thus, when we say that social norms are eroded, they do so according to either definition.

[2] This phenomenon is known in social psychology as "pluralistic ignorance" (Katz, Allport, and Jenness 1931), where privately most people reject a view, but incorrectly believe that most other people accept it, and therefore end up acting accordingly. For example, in 1968 most White Americans substantially overestimated the support for racial segregation among other Whites (O'Gorman 1975). A related concept is "preference falsification" (Kuran 1995): people's stated, public preferences are influenced by social acceptability, and might be different from their true, private preferences. For example, American college graduates consistently understate their support for immigration restrictions when asked directly as compared to their preferences elicited in a less obtrusive way, which is consistent with preference falsification (Janus 2010).

[3] A different mechanism, whereby powerful individuals can change norms by refusing to honor the systems of punishments and rewards that sustain these norms, is documented by Greif (2006) in the context of Genoan merchants and by Richman (2017) in the context of modern diamond traders.

[4] We thus focus on the consequences of Trump's election rather than its causes or determinants. Relatedly, Müller and Schwarz (2019) investigate the role of social media in spreading anti-Muslim sentiment during Trump's 2016 presidential campaign, while Giani and Méon (2019) document the effect of Trump's election on racial bias in policy attitudes outside of the United States. With respect to the determinants, Enke (forthcoming) demonstrates the link between communal (as opposed to universal) moral values and Trump vote at the county level, while Allcott and Gentzkow (2017) discuss the possible role of fake news. Relatedly, Xiong (forthcoming) studies the effect of the celebrity status of Ronald Reagan on his electoral support, and suggests that a similar effect may have helped Trump.

individuals likely disagreeing with xenophobic behavior sanction public expression of xenophobia less when they learn that the underlying view is more popular by observing the election outcome. This paper therefore studies a *natural* process of social norm *change*, and it does so by examining both agents and observers.

To organize thoughts and connect the experiments within a single framework, we build a model where two types of individuals, say xenophobic and tolerant, choose an action, but in doing so they care about approval or disapproval of other people who might observe the action (the relative proportion of the two types is a random variable whose realization is unknown to individuals). Like the agents choosing the action, the members of the audience are Bayesian, and their inference about the agent's type depends on the strategies he uses in equilibrium. In this environment, social pressure might lead some agents to choose the action that they do not naturally prefer, and it is even possible that all agents choose the same action (which, arguably, prevents learning about the distribution of types).[5] We then study the impact of public signals and show that a signal suggesting that more people are likely to be xenophobic increases the share of agents who choose the xenophobic action. However, the same signal may decrease the audience's perception that an agent who chose the xenophobic action is a xenophobe. Indeed, when few individuals are perceived to be xenophobic, there is no social pressure to appear to be one, and thus only xenophobes would choose a public xenophobic action. In contrast, when xenophobic individuals are thought to be common, such social pressure might be there, and thus not everyone acting in a xenophobic way is a true xenophobe.[6]

We experimentally capture the effects of Trump's rise in popularity. Throughout his campaign, Donald Trump proposed, among other things, the construction of a wall separating the United States and Mexico and a ban on Muslims from entering the United States. His popularity might thus send an informative signal about the number of people who sympathize with these proposals and thus about those who hold xenophobic views. As a result, Donald Trump's electoral success potentially caused a shift in social norms regarding expressing views on immigrants. We first examine the effect of Trump's rise in popularity (and thus of information aggregation) on people's willingness to publicly express xenophobia. In August–October 2018, we recruited a sample of 1,600 participants through an online panel survey company. We manipulate respondents' perceptions of Trump's local level of popularity in the 2016 election by exploiting the fact in some areas of the United States where that election was close, the candidate who won the election at the county level was different from the winner at the metropolitan statistical area (MSA) level. The subjects of this experiment were all recruited from the Pittsburgh MSA. In the beginning of the experiment, participants were given three facts about the history and politics of this area; we randomize whether one of these facts was that Donald

[5] The model is thus similar to earlier work on social image concerns and social norms (see Bénabou and Tirole 2006, for a general framework on incentives and prosocial behavior; and Ali and Lin 2013, for a model where ethical agents vote because they want to, while opportunistic ones vote in order to appear ethical). The possibility of pooling and the resulting non-transmission of information is suggested in Morris (2001), where an advisor who is afraid of being perceived as biased ultimately avoids giving informative advice.

[6] A similar effect appears in Bénabou and Tirole (2006), where an explicit incentive to do a prosocial action enables selfish people to do so, thus reducing the signaling value of such action. Similarly, in Ali and Lin (2013), a higher propensity of opportunistic voters (those caring about social image but not social welfare) to vote makes other such voters less willing to turn out.

Trump won the election in "Pittsburgh's metropolitan area" or that Hillary Clinton won the election in "Pittsburgh's county." Participants were then offered a bonus cash reward if they authorized the researchers to make a donation to a strongly anti-immigration organization on their behalf. Accepting the offer is therefore a *profitable* xenophobic action.

We also randomly manipulated the participants' expectations of anonymity through a public condition that implies visibility by a relevant "peer group," i.e., other subjects in the same geographic area of the respondent. All participants were told that the donation decisions would be posted on a website to be shared with all the participants from their area. One-half of the participants were assured that their individual responses would be kept completely anonymous (the "private" condition). The other half, instead, were exposed to a treatment designed to make them think that the responses posted on the website would not be anonymous (the "public" condition).

In the *Clinton Won* treatment, participants in the public condition were significantly more likely to forgo the donation bonus payment than those in the private condition. This suggests the presence of social stigma associated with the action. However, we find that for the participants in the *Trump Won* treatment, who received information that positively updated their perceptions of Trump's popularity, the wedge in the likelihood of undertaking the xenophobic action in private and public disappeared. This difference with respect to the *Clinton Won* condition was driven entirely by an increase in the donation rate in the public condition, with no change in the private condition. Our results suggest that an increase in perceptions of Donald Trump's local popularity does not make these participants more xenophobic, but instead makes those who were already xenophobic more comfortable expressing their xenophobic views in public. In addition, consistent with an underlying mechanism of updates in perceptions about the local popularity of xenophobic views, we find that the *Trump Won* treatment shifts the distribution of participants' perceptions about the local popularity of those views to the right. We also use the *Trump Won* treatment as an instrument for the shift in perceptions, and show that participants' perceptions causally affect their donation in public, but not in private. Note that instead of directly manipulating people's beliefs about a social norm (e.g., by telling them the share of people who hold a certain view), we instead vary the perceived local popularity of Donald Trump, and document that this naturally leads to changes in behavior.

Overall, these results indicate that social norms can quickly shift as a result of private opinions being aggregated and diffused through elections. In an earlier, motivating experiment implemented on Amazon Mechanical Turk (*mTurk*), we took advantage of a unique opportunity and examined information aggregation in "real time" in the weeks just before and after the November 2016 US Presidential election. The results are consistent with the causal effect we document in the main experiment. Participants who expected their decision to donate to a xenophobic organization to be observed by a surveyor were significantly less likely to make the donation than those expecting anonymity. Increases in participants' perceptions of Trump's popularity (either through experimental variation or through the "natural experiment" of his victory) eliminated the wedge between private and public behavior. The design and results from the motivating experiment are presented in online Appendix Section D.

We then turn our attention to the study of how the 2016 election changed people's sanctioning of observed xenophobic expression. We again hired an online panel survey company to recruit, in November 2018, a sample of about 1,800 respondents who had previously self-identified to the firm as Democrats. The respondents were asked to play a dictator game in which they decided how to split $2 between themselves and another agent. Our goal was to evaluate how they punished someone for expressing xenophobia in private or in public, depending on their perceptions of the popularity of xenophobia in that person's area. In particular, we were interested in studying the sanctioning of xenophobic expression in a situation where people believe that xenophobia is so popular that there might be social pressure for tolerant individuals to choose an observable xenophobic action. We told a recruited participant (henceforth, *player 1*) that a participant from a previous study (*player 2*) donated money to a named xenophobic organization, and cross-randomized whether we told player 1 that (i) Trump or Clinton won the election in player 2's area; and (ii) player 2's donation was in private or publicly observable on a website shared with other participants from his area. Note that participants assigned as player 2 were real participants from the experiment described above and that we selected only participants who decided to make the donation to the xenophobic organization. Other design elements are important to emphasize. Player 2 did not know he was going to be part of a dictator game in the future and player 1 was informed about this fact. Also, player 1 did not know that player 2's area was Pittsburgh (doing otherwise would have potentially limited the size of the shifts in player 1's perceptions about the popularity of xenophobia and the extent of social pressure for tolerant individuals to act as xenophobes in player 2's area). Finally, player 1 was not told that player 2 had been offered an incentive to donate to the organization.

We find that the amounts given by player 1 participants to player 2 participants are similar when player 2's donation was in private, regardless of who won the election in player 2's area. This helps us deal with confounds, such as learning about local motives to privately oppose immigration (since shifting perceptions on who won the election in player 2's area could have changed the perception of what that area is). Player 1 participants also give a similar amount to player 2 when his decision was in public when told that Clinton won in player 2's area. According to our two-type framework, the logic is simple: there is no social pressure to act as a xenophobe in an area where Clinton won (the pressure is, if anything, in the opposite direction), and therefore it must be that player 2 is indeed a xenophobe, just like in the private cases, which also did not involve social considerations. We do not find evidence, instead, of a different reaction from players 1 in the *Clinton Won* treatment, where players 2 who donate in public are sanctioned more heavily than those who donate in private, because by publicly expressing xenophobic views in an environment where those views are unpopular, players 2 reveal to be even more extreme than those who donate only in private (as could emerge from a framework with more than two types). In the final treatment, where player 2's decisions were public in an area where Trump won the election, player 1 participants give significantly more to player 2. Player 2 was now potentially subject to social pressure to publicly act as a xenophobe since his donation decision was going to be observed by local peers in an area where Trump won the election. The inference player 1 could make on whether player 2 was truly a xenophobe is therefore weakened due to the

strategic incentives player 2 could have to donate in order to pander to the majority. Here again, we use the information on who won the election in player 2's area as an instrument for player 1's perception of the popularity of xenophobia in that area, and show that these perceptions affect the amount given by player 1 when player 2's decision was public but not when it was private.

Our results suggest that Trump's rise in popularity and eventual electoral victory causally changed social norms regarding the expression of xenophobic views in the United States. Though we detect no changes in *privately held* views, we believe the findings on public expression are of great policy relevance.[7]

Related Literature.—Our results contribute to a growing literature that examines the impacts of political institutions on social norms and culture more generally. This literature typically studies the long-run impact of political institutions (e.g., Lowes et al. 2017); we show that changes on the political side can lead to fast changes in social norms. Our paper also adds to a recent theoretical literature on social norms (e.g., Bénabou and Tirole 2011, Acemoglu and Jackson 2017) by studying how new information may lead to erosion of such norms. Our findings also speak to a cross-disciplinary literature on the consequences of political actions, both theoretical (e.g., Lohmann 1993) and empirical (e.g., Madestam et al. 2013).[8]

Our work also relates to existing papers studying the economic consequences of conformity. Prendergast (1993) identifies rational incentives for managers to conform to supervisors' opinions in order to appear competent, which in turn hampers information transmission, while Bernheim (1994) shows that social concerns can lead to formation of social norms. Developing these ideas, Morris (2001) shows that the fear of being seen as biased could completely shut down information transmission from an advisor to a decision-maker. Ali and Lin (2013) study how social image concerns can give rise to prosocial behavior by non-altruistic individuals in the context of voter turnout, while Ali and Bénabou (2020) study how intensifying social image concerns about contributions to public goods makes it more difficult to learn about evolving social norms and preferences on those contributions.[9] Andreoni, Nikiforakis, and Siegenthaler (2017) study "conformity traps," situations where groups of individuals fail to coordinate on a beneficial action due to individual

[7] For example, increases in public expression of anti-immigrant sentiment might also lead to more frequent acts of hate crime against immigrants, and might also facilitate coordination for large-scale actions, such as demonstrations and movements. Recent work provides evidence that such demonstrations and movements might affect many important outcomes, from election results (Madestam et al. 2013) to the stock market valuation of different firms (Acemoglu, Hassan, and Tahoun 2018). In addition, a reduction in the stigma associated with holding previously extreme views might lead to shifts in the language used in and reported by the popular media, and might also reduce the stigma associated with consuming and discussing certain news sources on the far side of the political spectrum. An increase in public expression of such views can thus lead to an increase in individuals' overall exposure to them, and more exposure might eventually lead to changes in privately held views, via persuasion or simple conformism.

[8] Recent work has documented how policy changes, such as the introduction of cable TV in India (Jensen and Oster 2009) and quotas for female politicians (Beaman et al. 2009), can lead to changes in attitudes toward women and in behavior. However, these studies do not focus on isolating the mechanisms of social norm change (i.e., updates in perceived social acceptability as opposed to potential changes in private views.)

[9] Our model builds on these important precursors, albeit with notable differences. In these models, like in our paper, increased propensity of non-altruistic individuals to act prosocially makes it more difficult for observers to identify truly prosocial individuals. The main difference is that we seek to capture a contentious political setting where the same social image (e.g., of a xenophobic person) might be approved by some and stigmatized by others, as opposed to situations where all individuals seek to espouse a particular prosocial image.

incentives to conform to the predominant and inefficient behavior. In a laboratory experiment they find, in particular, that opinion polls can facilitate changes of norms that benefit the group. Their setting, however, is one of full information, and thus opinion polls facilitate switching from one equilibrium to another. Our model has incomplete information and features a unique equilibrium, and opinion polls or elections can change the beliefs about the distribution of other people's opinions. We also do not take a position on whether overcoming conformity is necessarily socially beneficial.

Our paper also contributes to a recent experimental literature on the effect of social norms on behavior. Krupka and Weber (2013) show that elicited social norms predict changes in behavior across variants of the dictator game. Bursztyn, González, and Yanagizawa-Drott (2020) directly manipulate perceived social norms, which in turn changes behavior. Here, our interest is in evaluating how natural processes that aggregate information about private opinions (such as an election) can lead to fast changes in social norms which lead both to changes in behavior *and* in sanctioning of behavior. Our approach also allows us to evaluate, using revealed preference, how updates about existing norms change both the way people express themselves and the way people sanction others for expressing themselves.

The remainder of this paper proceeds as follows. We introduce a simple framework formalizing our argument in Section I. In Section II, we present the design and results from Experiment 1, which studies the expression of xenophobic views. In Section III, we present the design and results from Experiment 2, which studies the sanctioning of xenophobic expression. Section IV concludes.

I. Motivating Framework

To organize thoughts and motivate our experimental designs, we present a simple model of social image and communication.

A. *Model*

The society consists of citizens of two types, with $\Pr\{t_i = A\} = p$ and $\Pr\{t_i = B\} = 1 - p$. This probability p is itself a random variable: $p = p_H$ with probability θ and $p = p_L$ with probability $1 - \theta$, where $\theta \in (0,1)$ and $0 < p_L < p_H < 1$. Conditional on the realization of p, the types of citizens are independent.

Each citizen knows his/her own type but not the realization of p. However, they get a public signal s that is informative of p. We assume that $s \in \{p_L, p_H\}$, and that $s = p$ with probability $\mu \geq 1/2$, so $\mu = 1/2$ corresponds to an uninformative signal, and $\mu = 1$ to a precise revelation of p.

Consider a citizen deciding between two actions, which we also denote A and B, slightly abusing notation (we will refer to the citizen choosing an action, the actor, as "he" and to a generic member of the audience as "she"). We interpret action A as the preferred action of type A and B as the preferred action of type B. More specifically, assume that the utility levels of each type from the corresponding action are $V_A > 0$ and $V_B > 0$, respectively, and the utility levels from the opposite actions are normalized to zero.

Suppose that citizen i is the actor choosing $d_i \in \{A, B\}$, and suppose he is doing so before an audience of size $n \geq 0$, with $n_{priv} = 0$ corresponding to a private decision and $n_{pub} > 0$ to a public decision (in which case let N_i denote this audience). Members of the audience observe the decision d_i and use it, as well as any other information they have, including the realization of their own type, to update their beliefs on citizen i's type t_i. We assume that citizen i gets utility χ_i if type A member of the audience believes he is type A (as opposed to B) or if type B one believes he is type B (as opposed to A).[10] Let $h_i = n_{pub}\chi_i$ denote individual i's total social image concerns when making a public decision and let j be a generic member of the audience; then citizen i makes decision d_i to maximize

$$(1) \quad U_i(d_i) = V_{t_i}\mathbf{1}\{d_i = t_i\} + \mathbf{1}_{pub}\chi_i \sum_{j \in N_i} E_i\left(\Pr_j(t_i = t_j \mid s, t_j, d_i) \mid s, t_i\right)$$

$$= V_{t_i}\mathbf{1}\{d_i = t_i\} + \mathbf{1}_{pub} h_i \sum_{t \in \{A,B\}} \Pr_i(t_j = s, t \mid t_i) \Pr_j(t_i = t \mid s, t_j = t, d_i).$$

Here, the last term reflects that both the sender and the receiver update on the distribution of types taking their own type into account: $\Pr_i(t_j = t \mid s, t_i)$ is the sender's posterior that receiver j has type t given public signal s and his type t_i, and $\Pr_j(t_i = t \mid s, t_j = t, d_i)$ is the receiver's posterior that sender has type t conditional on the receiver's type and, of course, the public signal and the sender's action.

We are interested in Perfect Bayesian equilibria that satisfy the D1 criterion (Cho and Kreps 1987). Throughout the paper we assume that social image concerns h_i are distributed smoothly and that direct payoffs from the preferred action (V_A and V_B) are not too small relative to these social image concerns. Formally, we impose the following assumption, and in online Appendix Section A.2 we show that otherwise we might get multiple equilibria.

ASSUMPTION 1: *Social image concerns h_i are independent of type t_i and are distributed on $[m, M]$ with c.d.f. $F(h)$ and p.d.f. $f(h)$ such that $f(h)h(h - \tilde{V}) \leq \tilde{V}$, where $\tilde{V} = \min\{V_A, V_B\}$. In addition, $\tilde{V} \geq (p_H - p_L)M$.*

[10] This reputational payoff is exogenously given in the model, but one can easily see how it may be endogenized, and how it may even get an instrumental component. In online Appendix Section A.5, we model the behavior of a receiver who can share money with the sender in a dictator game, and argue that modeling receivers as having homophily in addition to altruism predicts that the sender's payoff is monotonically increasing in the receiver's posterior about the sender's type. Our modeling choice maps well into our second experiment, and is consistent with existent empirical evidence. For instance, Hoffman, McCabe, and Smith (1996) document that social distance decreases transfers in dictator games, and Fowler and Kam (2007) show that donations to individuals of a different political orientation in a dictator game are lower. More broadly, perception of proximity of preferences between the two parties could have other material consequences, for example, by making credible communication possible in a Crawford and Sobel (1982) style model.

B. *Analysis*

Denote the share of type A in the society from the standpoint of a citizen who learned that his/her type is t and got public signal s by $P_t(s)$. By Bayes' formula, we have

$$P_A(s) = \frac{\theta(s)p_H^2 + \big(1-\theta(s)\big)p_L^2}{\theta(s)p_H + \big(1-\theta(s)\big)p_L};$$

$$P_B(s) = \frac{\theta(s)p_H(1-p_H) + \big(1-\theta(s)\big)p_L(1-p_L)}{\theta(s)(1-p_H) + \big(1-\theta(s)\big)(1-p_L)},$$

where

$$\theta(s) = \begin{cases} \dfrac{\theta\mu}{\theta\mu + (1-\theta)(1-\mu)} & \text{if } s = p_H \\ \dfrac{\theta(1-\mu)}{\theta(1-\mu) + (1-\theta)\mu} & \text{if } s = p_L \end{cases}$$

is the probability that $p = p_H$ conditional on signal s only. We will drop the argument s and write P_A and P_B if this does not cause confusion.

PROPOSITION 1: *Under Assumption 1 there is a unique Perfect Bayesian equilibrium that satisfies the D1 criterion. In the private decision case, citizen i chooses the action that corresponds to his type, $d_i = t_i$. In the public decision case:*

(*i*) *If $V_B \le mP_B(P_A + P_B - 1)$, then citizen i chooses $d_i = A$;*

(*ii*) *If $V_B \in \big(mP_B(P_A + P_B - 1), M(2P_B - 1)\big)$, then citizen i of type A chooses $d_i = A$, while citizen of type B chooses $d_i = A$ if h_i is above some threshold and $d_i = B$ otherwise;*

(*iii*) *If $V_B \ge M(2P_B - 1)$ and $V_A \ge M(1 - 2P_A)$, then citizen i chooses $d_i = t_i$;*

(*iv*) *If $V_A \in \big(m(1-P_A)(1-P_A-P_B), M(1-2P_A)\big)$, then citizen i of type B chooses $d_i = B$, while citizen of type A chooses $d_i = B$ if h_i is above some threshold and $d_i = A$ otherwise;*

(*v*) *If $V_A \le m(1-P_A)(1-P_A-P_B)$, then citizen i chooses $d_i = B$.*

These cases are exhaustive and mutually exclusive.

Let us now study comparative statics with respect to s. Notice that s enters utility (1) twice. The first instance, in $\Pr_i(t_j = t \mid s, t_i)$, is due to the actor believing that

the audience is more likely to consist of A types (B types) if $s = p_H$ ($s = p_L$). The second, in $\Pr_j(t_i = t \mid s, t_j = t, d_i)$, captures that the audience members' posterior that the actor is type A (B) is higher if $s = p_H$ ($s = p_L$). Both effects push the citizen toward choosing action A (B) if $s = p_H$ ($s = p_L$). The former effect is straightforward: if the audience is more likely to consist of A types, then choosing action A is more likely to boost one's social image from their viewpoint. The latter effect is slightly subtler: if a member of the audience has a strong prior that the sender is type A, it is difficult for a sender of type A to persuade her otherwise, and he might as well give up on the social image concerns and choose the action he likes.

The intuition above is summarized in the following proposition.

PROPOSITION 2: *In the public decision case, citizen is (weakly) more likely to choose $d_i = A$ (respectively, $d_i = B$) if he receives signal $s = p_H$ (respectively, $s = p_L$) as compared to the case of no signal.*

Let us now analyze how signal s affects the posterior probability that the citizen who chose action A is indeed type A. Recall that $P_B(p_L)$ denotes the posterior of type B individual about the share of type A in the society if the public signal is $s = p_L$.

PROPOSITION 3: *Suppose that $V_B > M(2P_B(p_L) - 1)$, so for low signal $s = p_L$, citizens of type $t_i = B$ choose action $d_i = B$. Then the audience's posterior that a citizen who chose action A is indeed type A is (weakly) lower for high signal $s = p_H$ than for low signal $s = p_L$, and strictly so if $V_B < M(2P_B(p_H) - 1)$. (The converse is true for the posterior probability that the citizen is type B.)*

In other words, if for signal $s = p_L$, indicating that type B is relatively frequent, citizens of type B choose the corresponding action, then anyone who chose action A must be indeed type A, i.e., the posterior is equal to 1. A signal $s = p_H$ indicating that type A is more frequent might lead to some B type citizens picking action A, which makes it possible that the citizen who chose action A is actually type B, which brings the posterior below 1. Of course, in the absence of an audience, i.e., in the private case, this posterior equals 1 for either signal. This posterior also equals to 1 in the public case with signal $s = p_L$, so, the public case with $s = p_H$ is the only one in which this posterior may be lower than 1. These comparative statics results are illustrated in online Appendix Section A.3 with a numerical example.

II. Experiment 1: Expressing Xenophobia

A. *Experimental Design*

Between August and October 2018, an online panel survey company recruited participants ($N = 1{,}600$) from the seven counties composing the Pittsburgh, PA metropolitan statistical area (Allegheny, Armstrong, Beaver, Butler, Fayette,

Washington, Westmoreland).[11] Each panelist could participate in the survey only once.[12]

After participants answered a number of demographic questions, we randomized the perceptions of Trump's popularity among the respondents' peers using a novel strategy that exploited the variation in the identity of the winner of the popular vote in the 2016 Presidential election across overlapping geographical areas. While Donald Trump won the popular vote in Pittsburgh's metropolitan area, Hillary Clinton won it in Allegheny, the county where Pittsburgh is located. At the beginning of the survey, respondents were told three facts about Pittsburgh history and politics. One-half of the participants were randomized into the *Trump Won* condition, and had one of the facts state that:

> In the 2016 US Presidential Election, Donald Trump won Pittsburgh's metropolitan area.

The other half of the participants were instead randomized into the *Clinton Won* condition, and were told that:

> In the 2016 US Presidential Election, Hillary Clinton won Pittsburgh's county.

The remaining two bullet points were the same for both groups of participants and contained neutral historical information about Pittsburgh politics. These two extra bullets were included to limit the participants' ability to infer the purpose of the study.

In the next part of the intervention, we measured the perceived social acceptability of strong anti-immigrant sentiment using a donation experiment with real stakes. Participants were first told that they would be given the opportunity to make a donation to a randomly drawn organization that could either be anti- or pro-immigration. Before the organization was revealed, the participants were also told that the donation would not be subtracted from their payment for participating in the survey, and that in case they authorized the donation they would be paid an extra $1. We analyze the behavior of the participants who were asked to donate to the organization we were interested in: the *Federation of American Immigration Reform*.[13] To make sure that the participants were aware of the organization's very strong anti-immigration

[11] The company, Qualtrics, is also used by Elías, Lacetera, and Macis (2019) and works as a panel aggregator that leverages a variety of actively managed participant databases. The databases consist of online participants who opted into participating in market research studies in exchange for incentives. On a quarterly basis, quality evaluations are performed by third-party data specialists to evaluate the databases on key aspects including response quality and consistency of the panelists.

[12] See the survey script in online Appendix C.2. The reader can access the survey information sheet and interactively follow the survey at http://ssd.az1.qualtrics.com/jfe/form/SV_erhJZh0coOf0u6F (accessed on July 6, 2020).

[13] Our original goal was to assign only 1 percent of participants to make a donation to the pro-immigration organization, since this decision was not our outcome of interest, and we wanted instead to maximize power for the analysis of donations to the anti-immigration organization. In other words, we focused on xenophobic behavior, and therefore did not test the full set of predictions of our model, in particular, the stigmatization of tolerant behavior. As a result, we did not plan to have enough observations (and thus statistical power) to analyze donations to the pro-immigration organization. Unfortunately, due to an implementation error, 100 percent of the participants were actually assigned to the anti-immigration organization.

stance, a few more details about the organization and its founder were provided in the experiment:

> The Federation for American Immigration Reform (FAIR) is an *immigration-reduction organization* of concerned individuals who believe that immigration laws must be reformed, and seeks to reduce overall immigration (both legal and illegal) into the United States. The founder of FAIR is John Tanton, author of "The Immigration Invasion" who wrote "I've come to the point of view that for European-American society and culture to persist requires a European-American majority, and a clear one at that."

In addition to the first randomization informing subjects about either Trump's or Clinton's victory, we introduced a second layer of cross-randomization at the donation stage. All of the participants were told that the results of the survey would be posted online. The link to the website was given to all participants, so that it was clear to each participant that other participants (who came from the same geographical area around Pittsburgh, a fact known to the participants) would have access to the result of the survey. Importantly, however, one-half of the participants were assured that their individual responses would be kept completely anonymous, and that the results would be posted in aggregate form as percentages only: we refer to this condition as the *private* condition. Specifically, participants were told:

> The anonymized results from this survey will be posted on our website in approximately one month. Results will be reported as percentages only, so your individual response to the survey will remain anonymous. We will notify you when the results become available on our website, [website link].

The other half of the subjects were not given this assurance of anonymity. In practice, their decision appeared on the website in anonymous form as for the other half of the participants. Importantly, however, their decisions were not reported as percentages in aggregated form: instead, for each participant, the website reported the (anonymous) survey response ID and the individual donation decision. To avoid deception, the subjects were never explicitly told that their personal details would be published on the website along with their donation decision. The participants, however, were given clues suggesting that their name and email could be published on the website together with their individual donation decision. This is what we refer to as the *public* condition:

> The results from this survey, including your individual donation decision and the donation decisions of all of the other Pittsburgh respondents to this survey, will be posted on our website in approximately one month. There is no need to provide your name, email, etc. here; the survey company we work with has this information already. We will notify you when the results become available on our website, [website link].

Note that the survey company indeed had access to the participants' personally identifiable information, but we (the experimenters) did not. As a result, the statements in both conditions were factually true.

After the donation decision, participants were asked to predict the share of Pittsburgh voters who agree with the following anti-immigration statements:

> For European-American society and culture to persist requires a European-American majority, and a clear one at that.

and

> Both legal and illegal immigration should be drastically reduced because immigrants undermine American culture and do not respect American values.

This provides two measures of the perceived local popularity of anti-immigrant sentiment. At the end of the survey, the respondents answered a few demographic questions.

B. *Link to Theory*

This experiment looks at the sender's decision. We interpret type A as xenophobic and type B as tolerant; action A as the xenophobic action (authorizing the donation to the anti-immigration organization) and action B as the tolerant one (not authorizing this donation). Absent social image concerns, xenophobic individuals should strictly prefer action A (they help an aligned organization and get a dollar) and tolerant ones should strictly prefer action B (implicitly, we assume that associating with the organization creates more than a dollar of disutility for tolerant people). The citizens are our subjects (survey participants), and the audience (in the public setting) are those who would visit the website we provided. We interpret information that Trump won Pittsburgh MSA as the high signal about the share of type A, $s = p_H$, and information that Clinton won Pittsburgh's county as the low signal $s = p_L$.

In terms of first differences, the model (Proposition 2) predicts that citizens who got signal $s = p_H$ are weakly more likely to choose the xenophobic action than those who got signal $s = p_L$ in the public setting, and that there is no difference in the private setting. Moreover, within the $s = p_L$ setting, people are weakly less likely to choose the xenophobic action in public than in private, and the opposite is true if $s = p_H$.

The extent to which these inequalities are strict depend on the parameters of the model (as Proposition 1 helps clarify). If $s = p_L$, citizens are strictly less likely to choose the xenophobic action in public than in private if $V_A < M\big(1 - 2P_A(p_L)\big)$, that is, if the direct payoff for a xenophobe from choosing the xenophobic action V_A is lower than the social pressure to choose the tolerant action because his posterior about the share of xenophobes $P_A(p_L)$ is low. For signal $s = p_H$, however, citizens are strictly more likely to choose the xenophobic action in public than in private if $V_B < M\big(1 - 2P_B(p_H)\big)$, that is, if the direct payoff for a tolerant from choosing the tolerant action V_B is lower than the social pressure to choose the xenophobic action because his posterior about the share of xenophobes $P_B(p_H)$ is high. Finally, citizens who got signal $s = p_H$ are strictly more likely to choose the xenophobic

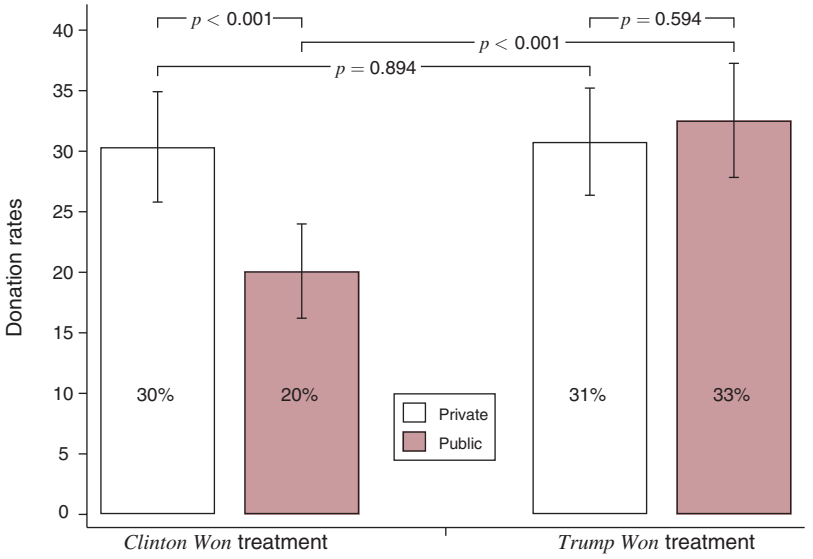

FIGURE 1. EXPERIMENT 1: DONATION RATES

Notes: The two bars on the left display donation rates to the anti-immigration organization for individuals in the private and public conditions in the *Clinton Won* treatment group (respectively $N = 392$ and $N = 408$), and the two bars on the right display those for individuals in the private and public conditions in the *Trump Won* treatment group (respectively $N = 419$ and $N = 381$). Error bars reflect 95 percent confidence intervals. Top horizontal bars show p-values for t-tests of equality of means between different experimental conditions.

action than those who got signal $s = p_L$ in the public setting as long as at least one of the two conditions are satisfied.[14]

C. *Main Results*

Online Appendix Table B1 provides evidence that individual characteristics are balanced across all four experimental conditions, confirming that the randomization was successful. Figure 1 displays the main findings from this experiment. The experimental results support the predictions of the model.[15] First, public donation rates are higher in the *Trump Won* condition than in the *Clinton Won* condition (33 percent and 20 percent, respectively, p-value < 0.001). Second, we find no significant difference in private donation rates between the *Trump Won* and the *Clinton Won* conditions (respectively, 31 percent and 30 percent with p-value $= 0.894$), suggesting

[14] One caveat is that the model assumes that the same signal is obtained by both the senders and receivers, which corresponds to survey participants expecting the audience to consist of other survey participants or those who knew which politician won the election in Pittsburgh area. This is reasonable, as we provided the link to the survey participants only. However, if they expected the results of the experiment to be available to a larger internet audience, including those who did not know who won in the Pittsburgh metropolitan area or county but who nevertheless mattered for the purposes of social image concerns, this would require a model where senders get the signal but receivers do not. As we show in online Appendix Section A.4, the results of Proposition 2 continue to hold as stated. Thus, the predictions of our framework are robust to our subjects' interpretation of who exactly will get access to the results of the survey.

[15] Specifically, for parameter values where $V_A < M\big(1 - 2P_A(p_L)\big)$, but $V_B \geq M\big(2P_B(p_H) - 1\big)$. In other words, the results are consistent with subjects believing that there are sufficiently few xenophobes in Pittsburgh after getting $s = p_L$, and not too many even if $s = p_H$.

that the information about the winner of the popular vote is not changing privately held xenophobia. The difference in differences between donation rates in private across conditions and donation rates in public across conditions is statistically significant (p-value $= 0.008$). Moreover, we find that these results are primarily due to a significant wedge between private and public donations rates in the *Clinton Won* (the p-value of a t-test of equality is <0.001), with no significant difference in the *Trump Won* treatment (p-value $= 0.594$). As discussed above, these results are also consistent with the model, and indicate that in a city like Pittsburgh the victory of Trump did not result in a significant pressure for tolerant individuals to behave as xenophobes in public, while knowledge about Clinton victory is associated with social pressure for xenophobic individuals to appear more tolerant in public. To reiterate, these results indicate that the information provided about Trump winning the popular vote in the Pittsburgh area causally increased the social acceptability of the action to the point of eliminating the social stigma associated with it among the participants who were told that Clinton won the popular vote.[16] Table 1 displays the difference in differences results in regression format and shows that our results are unchanged when individual covariates are included.[17] The table also displays p-values from permutation tests, showing that our findings are robust to that inference method.

D. *Evidence of Mechanism*

Following the donation decisions, we also asked participants about their perceptions of the local popularity of anti-immigrant sentiments. We note that asking these questions after the donations could lead to mismeasurement of perceptions if, for example, participants inflated their reported perceptions of anti-immigrant sentiments because they just made an anti-immigration donation themselves. Because of this, the analysis in this section should be taken with caution. Nevertheless, the results presented here provide useful insights for the understanding of the possible mechanisms behind our main findings, and are consistent with the interpretation that the electoral victory of Donald Trump increased the willingness of participants to express xenophobic views through an update in their perceptions of the popularity of anti-immigrant sentiments.[18]

[16] It is important to note that Experiment 1 identifies the relative effect of information about the two candidates winning the local popular vote, but is not informative about whether it was the victory of Trump that increased the social acceptability of xenophobic actions, or if instead it was the victory of Clinton that could have generated social stigma for acting in a xenophobic way. The results from Experiment 1B, presented in the online Appendix, suggest that the former hypothesis holds, i.e., that there was a "Trump effect" rather than a "Clinton effect." We document that, before the election, participants in the control condition who received no information about the rising popularity of Trump were less likely to make a donation to the anti-immigrant organization in public than they were in private. However, increases in participants' perceptions of Trump's popularity (either through experimental variation or through the "natural experiment" of his victory) eliminated the wedge between private and public behavior.

[17] Heterogeneous treatment effects by race, gender, age, marital status, education, and income are reported in online Appendix Table B2. The direction of the treatment effect is the same in all subgroups, and differences in the magnitude of the effects between subgroups are never statistically significant. Point estimates indicate, however, that the wedge in donations between public and private in the *Clinton Won* treatment is lower among Whites than among non-Whites, and that the reduction in the wedge in the Public condition is complete among Whites but only partial among non-Whites.

[18] We ask these questions after the donation, and not before, because in this study we are interested in understanding whether information from elections per se leads to a *natural* change in social norms, without any need for external additional information priming people to think about the popularity of those norms. Between having

Table 1—Experiment 1: Difference-in-Differences Regressions

	Dummy: individual authorizes donation to anti-immigration organization	
	(1)	(2)
Public	−0.103	−0.109
	(0.031)	(0.031)
	[0.001]	[<0.001]
Trump Won	0.004	0.000
	(0.032)	(0.033)
	[0.897]	[0.988]
Public × *Trump Won*	0.120	0.130
	(0.045)	(0.045)
	[0.002]	[0.001]
Mean donation rate *Clinton Won* private treatment	0.304	
Controls	No	Yes
Observations	1,600	1,587
R^2	0.012	0.022

Notes: Column 1 presents OLS regression of a dummy variable for whether an individual donates to the anti-immigration organization on a dummy for the public condition, a dummy for the *Trump Won* condition, and a dummy for the *Trump Won, public* condition. The *Clinton Won, private* condition is the omitted group, for which we report the mean donation rate. Column 2 replicates and adds individual covariates (gender, age, marital status, years of education, household income, and race). Robust standard errors in parentheses. *p*-values from permutation tests with 1,000 repetitions in brackets.

Figure 2 shows that, consistent with the underlying mechanism of updates in perceptions about the local popularity of xenophobic views, the distribution of perceptions about the local popularity of these views held by participants in the *Trump Won* treatment is to the right of the distribution of perceptions held by participants in the *Clinton Won* treatment. The average beliefs are, respectively, 50.1 percent and 42.6 percent (p-value < 0.001).

Given this strong difference in perceptions caused by the treatment, we also use the *Trump Won* treatment as an instrument for perceptions. In Table 2, we present evidence consistent with the idea that participants' perceptions causally affect their donation in public, but not in private. Panel A replicates the results of Figure 2, and additionally shows that the *Trump Won* treatment is associated with higher perceptions of the popularity of anti-immigrant views both in the private and in the public conditions. Moreover, these results are robust to the inclusion of individual covariates. In panel B, we show that a one percentage point increase in the perception of the local popularity of xenophobia increases donation rates in public by 2.3 percentage points. On the contrary, as expected, donations in private are unaffected by the experimentally induced difference in beliefs.[19]

this elicitation being potentially affected by our main outcome (the donation) or having our main outcome being potentially affected by the elicitation, we chose the former since we wanted to avoid priming our respondents before they made their donation decision.

[19] Beyond the caveat on the measurement of perceptions discussed above, we view the IV results as suggestive also because it is possible that the *Trump Won* treatment affects donation rates through other channels. However, we find this unlikely. Indeed, all alternative channels not associated with social acceptability are ruled out by the

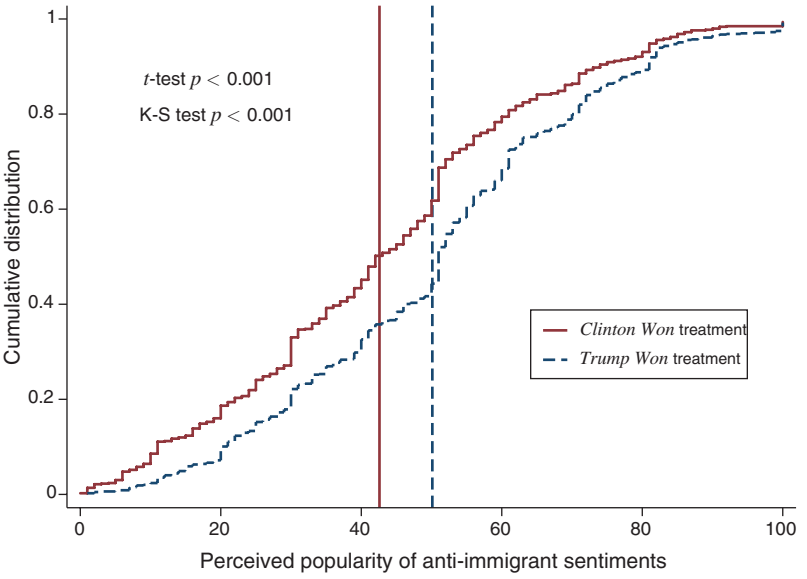

FIGURE 2. EXPERIMENT 1: BELIEFS ABOUT OTHERS

Notes: Empirical cumulative distributions of perceived popularity of anti-immigrant sentiments for individuals in the *Clinton Won* treatment and in the *Trump Won* treatment (respectively $N = 793$ and $N = 794$). The two vertical lines display the means of the two distributions. K-S p is the p-value of a Kolmogorov-Smirnov test of equality of the two distributions, while t-test p is the p-value of a test of equality of means.

III. Experiment 2: Sanctioning Xenophobic Expression

A. *Experimental Design*

In November 2018, an online panel survey company recruited participants ($N = 1{,}830$) from the United States who had previously self-identified as Democrats.[20] Each panelist could participate in the survey only once. The survey was conducted on the online platform Qualtrics.

First, all participants answered a number of demographic questions. The participants of this experiment were randomized into one of four conditions, corresponding to the four treatments of Experiment 1: the *Clinton Won*, *private* group,

fact that the donations in private do not change. A remaining possibility is that the *Trump Won* treatment might have changed the perceived local acceptability of other behaviors, for example of acting selfishly: after all, Donald Trump could be seen both as the anti-immigrant candidate and as the candidate of greed and self-interest (we thank an anonymous referee for raising this point). The results of Experiment 3 (reported in online Appendix F), however, help us rule out this hypothesis. The design of Experiment 3 is similar to that of Experiment 1: it uses donation decisions made either in a private or in a public condition to study the social acceptability of a view. In Experiment 3, however, instead of varying the perceived local popularity of candidate Trump as we do in Experiment 1, we directly randomize the perceived local popularity of anti-Muslim sentiments. Consistently with an update in the perceived popularity of the view being the mechanism at play in both Experiment 1 and 3, we find similar patterns in both experiments.

[20] The company we used for this experiment is Prime Panels. See the survey script in online Appendix Section C.3. The reader can access the survey information sheet and interactively follow the survey at http://ssd.az1.qualtrics.com/jfe/form/SV_4VgnEZSmikzSf8p (accessed on July 6, 2020).

Table 2—Experiment 1: Instrumental Variable Regressions

	Panel A. First-stage regressions Perceived share of voters holding xenophobic views					
	(1)	(2)	(3)	(4)	(5)	(6)
Trump Won treatment	7.528	7.568	9.492	9.384	5.541	5.668
	(1.111)	(1.101)	(1.568)	(1.548)	(1.580)	(1.574)
Mean perceived share *Clinton Won* treatment	42.588		41.451		43.666	
	Panel B. Instrumental variable regressions Dummy: individual authorizes donation to anti-immigration organization					
	(1)	(2)	(3)	(4)	(5)	(6)
Perceived share of voters holding xenophobic views			0.000	0.000	0.023	0.024
			(0.003)	(0.003)	(0.008)	(0.008)
Mean donation rate at a 50 percent perceived share of voters holding xenophobic views			0.305		0.344	
Sample	Full sample		Private treatment		Public treatment	
Controls	No	Yes	No	Yes	No	Yes
Observations	1,587	1,587	802	802	785	785

Notes: Panel A presents OLS regressions of the perceived share of voters holding xenophobic views on a dummy for the *Trump Won* treatment. Specifically, we measure the percentage of Pittsburgh voters the respondents believed would agree with the quote "For European-American society and culture to persist requires a European-American majority, and a clear one at that." The *Clinton Won* treatment is the omitted group, for which we report the mean perceived share. For columns 3, 4, 5, and 6 these estimates also represent the first stage of the instrumental variable regressions presented in panel B. Panel B presents IV regressions of a dummy variable for whether an individual donates to the anti-immigration organization on the perceived share of voters holding xenophobic views. The latter is instrumented with the *Trump Won* treatment. We subtract 50 from the measure of the share, so that the intercept of the regression represents the average donation rate at a perceived share of 50 percent. Columns 2 and 4 and 6 replicate and add individual covariates (gender, age, marital status, years of education, household income, and race). Robust standard errors in parentheses.

the *Clinton Won*, *public* group, the *Trump Won*, *private* group, and the *Trump Won*, *public* group. Those in the two *Clinton Won* groups were told that we surveyed, in another survey, participants from an area where Hillary Clinton won in 2016. Those in the two *Trump Won* groups were told instead that we surveyed participants from an area where Donald Trump won in 2016.[21]

In Experiment 1, we studied how varying social pressure against xenophobia affected xenophobic expression. In Experiment 2, we focus on how varying social pressure for xenophobic expression affects sanctioning of this type of expression. In particular, we were interested in studying the sanctioning of xenophobic expression in a situation where people believe that xenophobia is so popular that there might be enough social pressure for tolerant individuals to choose the observable xenophobic action. To that end, we made an ex ante design choice not to disclose to the participants of Experiment 2 that the previous study was about Pittsburgh. This would allow for a larger positive update in perceptions about the local popularity

[21] The information provided to both groups of participants, although similarly contradictory, is nonetheless factually true, with the caveat that it relies on two different definitions of "areas": the county for the *Clinton Won* condition, and the MSA for the *Trump Won* condition. Participants were not informed about the precise definition of "area" used in their treatment.

of xenophobia stemming from information about Trump's victory in the area of the first study (and it is indeed what we observe empirically, as shown in Figures 2 and 4, which display participants' beliefs for both experiments).[22]

All subjects were then presented with two anti-immigrant quotes (the same used in Experiment 1) and were asked to predict the share of voters in the other participant's area that they believed would agree with the quotes. This provides two measures of the beliefs that the participants in this experiment had about the popularity of anti-immigrant sentiments in the area where the previous study took place.

Next, the subjects (players 1) were informed that they had been matched with a participant from the previous survey (player 2). Each player 1 was matched with a random player 2 from one of the four original conditions. For example, a player 1 randomized into the *Clinton Won*, *private* group for Experiment 2 is defined as one who was matched with a player 2 that was randomized into the *Clinton Won*, *private* condition in Experiment 1. Players 1, however, were not informed that the previous survey was an experiment with multiple treatment conditions. The subjects were then told that player 2 authorized a donation to an anti-immigration organization, after being shown the exact text of the question in which the donation was authorized. Importantly, the text included either the text of the *private* treatment or of the *public* treatment, so that the subject could fully understand the condition under which the other participant made the donation choice. For example, players 1 in the *Trump Won*, *public* group knew that the player 2 they were matched with was from an area where Donald Trump won the 2016 election and decided to authorize a donation to the anti-immigration organization knowing that their donation decision would be published online, possibly along with their personal details. The subjects, however, were not informed about the fact that the donation in the previous survey was incentivized. [23]

Finally, players 1 were asked to play a dictator game in which they could decide how to split \$2 between themselves and player 2. The subjects were told that their decision on how much to give to the other participant was anonymous, and that when making his donation decision, player 2 did not know that they would be playing this follow-up game.

B. *Link to Theory*

This experiment looks at receivers' interpretation of senders' decisions. As above, we interpret type A as xenophobic and type B as tolerant; action A as the senders' xenophobic action (authorizing the donation to the anti-immigration organization) and action B as the tolerant one (not authorizing this donation). Absent social image concerns, xenophobic senders should have strictly preferred action A (they helped an aligned organization) and tolerant ones should have strictly preferred action B (they

[22] In other words, in Experiment 1 we found ourselves in the region of the model where $V_B \geq M\big(2P_B(p_H) - 1\big)$: there was social pressure for xenophobes to choose tolerant actions in the *Clinton Won* treatment. In Experiment 2, we chose the design to ensure that the *Trump Won* signal was strong enough to ensure that $V_B < M\big(2P_B(p_H) - 1\big)$, so there is nontrivial social pressure for tolerant individuals to act as xenophobes. In this case, Proposition 3 predicts a strict inequality in the posteriors following the two signals.

[23] This was done to facilitate player 1's inference about player 2's motivation to make the donation: we eliminate the possibility of an inference that player 2 made the donation because of financial incentives, that is, independently of private views or social pressure.

refused to associate with the organization). Moreover, the subjects of Experiment 2 (players 1 in the dictator game) are the receivers in the model, who judge the action of the senders and decide how much to share with them in a dictator game. All the senders (players 2) have chosen action A, since in Experiment 1 they decided to authorize the donation to the anti-immigration organization: their type, however, is not directly known, but can be inferred by the receivers. We think of participants as type-B members of the audience, since they are all Democrats (and thus unlikely to be xenophobic). Finally, we interpret information that Trump won as the high signal about the share of type A in the senders area, $s = p_H$, and information that Clinton won as the low signal $s = p_L$, for which we believe no type-B tolerant senders would take the xenophobic action A in public (that is, as in Proposition 3, we assume $V_B > M\big(2P_B(p_L) - 1\big)$ to rule out the case in which even the low signal is enough to generate social pressure for tolerant individuals to act as xenophobes).

The model predicts that in private only type-A senders will chose action A. As a result, the receivers' posterior that a player 2 who chose the xenophobic action is a xenophobe equals 1 in the private case, regardless of which candidate won locally. According to Proposition 3, in public, this posterior is also equal to 1 in the case of the low signal $s = p_L$ (there is no social pressure for someone tolerant to act as a xenophobe in the *Clinton Won*, *public* treatment). In contrast, with the high signal $s = p_H$, some tolerant senders might choose the xenophobic action due to social pressure. As a result, the posterior that player 2 who chose the xenophobic action is a xenophobe would be weakly less in the *Trump Won*, *public* treatment than in the *Clinton Won*, *public* treatment, and strictly so if $V_B < M\big(2P_B(p_H) - 1\big)$, i.e., if the posterior of tolerant people about the share of xenophobes following $s = p_H$ is sufficiently high.

Since all receivers (player 1) in our experiment are Democrats, the lower posterior that player 2 is a xenophobe should result in lower punishment/higher altruism toward player 1. Thus, we would expect donations by dictators to be higher in the *Trump Won*, *public* treatment than in any other treatment, and the donations in the other three treatments to be similar.[24]

It is important to emphasize that in a more general model that allows for more than two levels of xenophobia, player 1 could infer that a player who donated in the *Clinton Won*, *public* treatment is more xenophobic than one who donated in the *Clinton Won*, *private* treatment. Indeed, by publicly expressing xenophobic views in an environment where those views are unpopular player 2 would reveal to be even more extreme than when donating in private. In this case, the punishment by player 1 would be higher for a player 2 in that condition, something that cannot be the case in the present, two-type version of the model.

C. *Main Results*

Online Appendix Table B3 provides evidence that individual characteristics are balanced across all four experimental conditions, confirming that the randomization was successful.

[24] In online Appendix Section A.5 we provide a simple microfoundation for the link between posteriors about player 1 and donations to player 1 based on homophily and altruism, as discussed in footnote 10.

Figure 3 displays our main findings from Experiment 2. Panel A displays comparisons of average donations across groups. In the *Clinton Won*, *private*, *Clinton Won*, *public*, and *Trump Won*, *private* conditions the average amounts given are very similar, at \$0.78, \$0.81, and \$0.78, respectively. The *p*-value of a test of joint equality of the three averages is 0.739. In each of these conditions there was no pressure to donate to the anti-immigration organization in order to pander to the majority: in the two private conditions there is no social pressure, and in the *Clinton Won*, *public* condition the desire to pander to the majority would, if anything, pressure the subject into not authorizing the donation.[25] From the decision to donate, the subject could then infer that player 2 was likely to truly hold xenophobic views. The lack of difference across the two private conditions helps us deal with alternative interpretations, such as learning about local motives to privately oppose immigration (since shifting perceptions on who won the election in player 2's area could have changed the perception of what that area is).

In the *Trump Won*, *public* condition the average donation is \$0.88, higher than in the *Clinton Won*, *private* condition (p-value 0.006), the *Clinton Won*, *public* condition (p-value 0.051), and *Trump Won*, *private* condition (p-value 0.008). The one in the *Trump Won*, *public* condition is the only donation that could have been driven by the desire of player 2 to pander to the majority, complicating the inference that player 1 could make about the anti-immigration views of the participant from the previous study.

Panel B compares the share of participants who do *not* share anything from their \$2 endowment with player 2. Here again, the percentages of subject deciding not to transfer anything to the other participant are similar across the *Clinton Won*, *private*, *Clinton Won*, *public*, and *Trump Won*, *private* conditions, at respectively 18 percent, 21 percent, and 22 percent. The *p*-value of a test of joint equality of the three averages is 0.374. Importantly, the share of participants not donating is 8 percent in the *Trump Won*, *public* group, significantly lower than in the other three conditions (p-value < 0.001 for all three pairwise comparisons).[26]

Online Appendix Table B4 displays the results in regression format and shows that our results are not changed when individual covariates are included.

In online Appendix E we present the design and results from a similar experiment conducted on *mTurk*, where participants were asked to play a dictator game with another respondent in Switzerland, and where we manipulated perceptions of the popularity of anti-Muslim sentiment in Switzerland, by randomly giving information about the 2009 Swiss referendum that banned the construction of minarets in that country.

[25] Indeed, as suggested above, in a model with more than two types—for example, with tolerant, weakly xenophobic, and strongly xenophobic individuals—in the *Clinton Won*, *public* condition, players 1 could sanction more heavily those who donate in public than those who donate in private, because by publicly expressing xenophobic views in an environment where those views are unpopular, players 2 reveal to be even more extreme than those who donate only in private (for example, only strongly xenophobic individuals would donate in public, while both weakly and strongly xenophobic individuals would donate in private). We do not find evidence consistent with this alternative.

[26] The median amount given was \$1 in all four treatments, so we do not use it as an outcome.

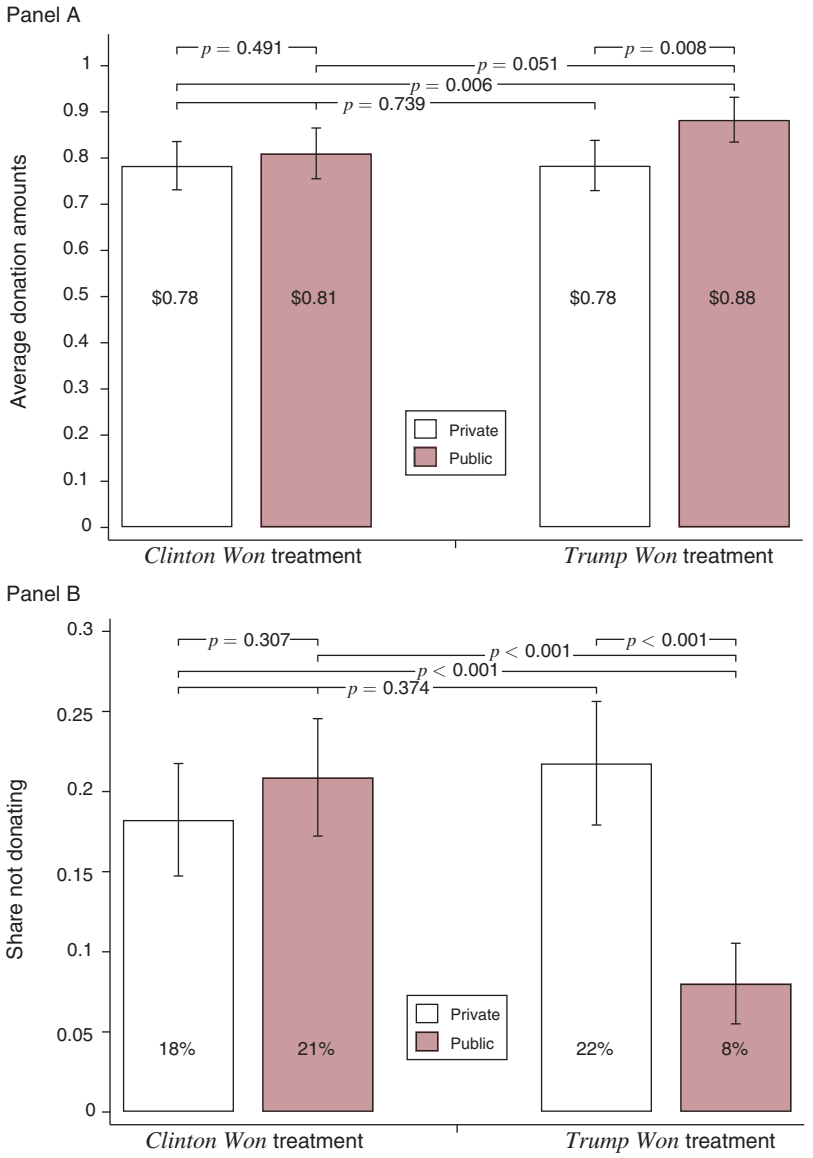

FIGURE 3. EXPERIMENT 2: DONATION RATES

Notes: Panel A displays average donation amounts to the individuals from Experiment 1 in the four experimental conditions. The two bars on the left display donations for individuals in the private and public conditions in the *Clinton Won* treatment group (respectively $N = 466$ and $N = 474$), and the two bars on the right display those for individuals in the private and public conditions in the *Trump Won* treatment group (respectively $N = 441$ and $N = 449$). Panel B displays the percent of subjects not making positive donations. Error bars reflect 95 percent confidence intervals. Top horizontal bars show p-values for t-tests of equality of means between different experimental conditions.

D. *Evidence of Mechanism*

Figure 4 shows that, consistent with the underlying mechanism of updates in perceptions about the local popularity of xenophobic views in player 2's area, the distribution of perceptions about the local popularity of these views held by dictators in

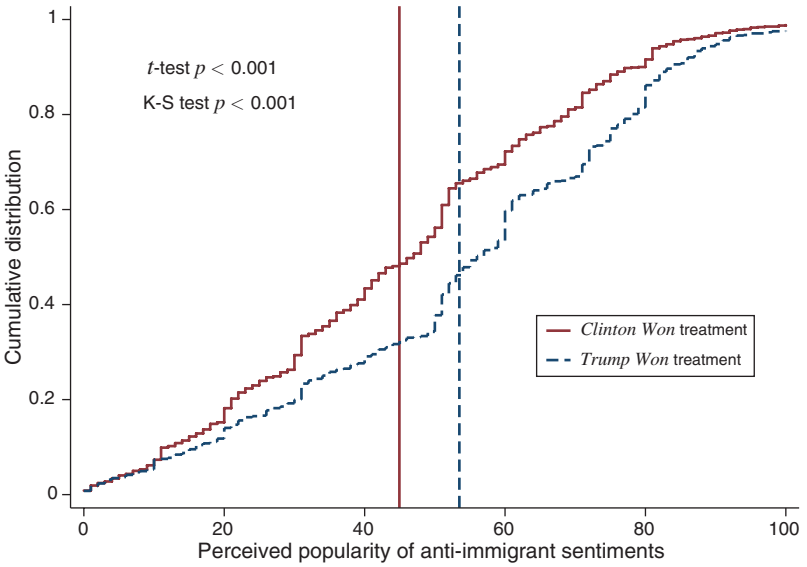

FIGURE 4. EXPERIMENT 2: BELIEFS ABOUT OTHERS

Notes: Empirical cumulative distributions of perceived popularity of anti-immigrant sentiments for individuals in the *Clinton Won* treatment and in the *Trump Won* treatment (respectively $N = 940$ and $N = 890$). The two vertical lines display the means of the two distributions. K-S test p is the p-value of a Kolmogorov-Smirnov test of equality of the two distributions, while t-test p is the p-value of a test of equality of means.

the *Trump Won* treatment is to the right of the distribution of perceptions held by dictators in the *Clinton Won* treatment. The average beliefs are respectively 53.5 percent and 44.9 percent (p-value < 0.001).

We can again use the *Trump Won* treatment as an instrument for perceptions. In Table 3, we present evidence that the dictators' perceptions causally affected the amount they gave and the probability of giving when the subject from Experiment 1 donated in public, but not when they donated in private. Panel A replicates the results of Figure 4, and additionally shows that the *Trump Won* treatment is associated with higher perceptions of the popularity of anti-immigrant views both in the private and in the public conditions. Moreover, these results are robust to the inclusion of individual covariates. In panel B, we show that a 1 percentage point increase in the perception of the local popularity of xenophobia increases the amount shared by dictators in the public condition by \$0.007 (the IV estimate however is noisy, with a p-value of 0.054). On the contrary, as expected, donations in the private conditions are unaffected by the experimentally induced difference in beliefs. In panel C, we show that a 1 percentage point increase in perception of the local popularity of xenophobic views is also associated with a 1.2 percentage points decrease in the share of dictators deciding not to transfer anything to players 2 who made a donation in the public condition (p-value < 0.001), while there are again no significant differences in the private condition. Taken together, these results indicate that manipulating beliefs about the relative popularity of Donald Trump: (i) has the expected effect on the perceived popularity of anti-immigrant views in the area of player 2; (ii) is associated with lower punishment in the public condition (which is

Table 3—Experiment 2: Instrumental Variable Regressions

	Panel A. First-stage regressions Perceived share of voters holding xenophobic views					
	(1)	(2)	(3)	(4)	(5)	(6)
Trump Won treatment	8.544	8.795	6.438	6.692	10.614	10.820
	(1.155)	(1.148)	(1.666)	(1.662)	(1.597)	(1.590)
Mean perceived share *Clinton Won* treatment	44.948		45.118		44.781	
	Panel B. Instrumental variable regressions Amount donated in the dictator game					
	(1)	(2)	(3)	(4)	(5)	(6)
Perceived share of voters holding xenophobic views			0.000	0.000	0.007	0.007
			(0.006)	(0.006)	(0.004)	(0.003)
Mean dependent variable at a 50 percent perceived share of voters holding xenophobic views			0.784		0.846	
	Panel C. Instrumental variable regressions Dummy: individual does not share anything in the dictator game					
	(1)	(2)	(3)	(4)	(5)	(6)
Perceived share of voters holding xenophobic views			0.005	0.005	−0.012	−0.012
			(0.005)	(0.004)	(0.003)	(0.003)
Mean dependent variable at a 50 percent perceived share of voters holding xenophobic views			0.209		0.146	
Sample	Full sample		Private treatment		Public treatment	
Controls	No	Yes	No	Yes	No	Yes
Observations	1,830	1,830	907	907	923	923

Notes: Panel A presents OLS regressions of the perceived share of voters holding xenophobic views on a dummy for the *Trump Won* treatment. Specifically, we measure the percentage of Pittsburgh voters the respondents believed would agree with the quote "For European-American society and culture to persist requires a European-American majority, and a clear one at that." The *Clinton Won* treatment is the omitted group, for which we report the mean perceived share. For columns 3, 4, 5, and 6 these estimates also represent the first stage of the instrumental variable regressions presented in panel B. Panel B presents IV regressions of a dummy variable for whether an individual donates to the anti-immigration organization on the perceived share of voters holding xenophobic views. The latter is instrumented with the *Trump Won* treatment. We subtract 50 from the measure of the share, so that the intercept of the regression represents the average donation rate at a perceived share of 50 percent. Columns 2 and 4 and 6 replicate and add individual covariates (gender, age, marital status, years of education, household income, and race). Robust standard errors in parentheses.

when the social pressure for the recipient to express those views could be present); and (iii) does not have any effect in the private condition (where social pressure is absent).

IV. Conclusion

In this paper, we study how social norms, usually thought of as relatively stable and persistent, can change rapidly when new information becomes available. In our first experiment, we show that a positive, experimentally induced update in people's beliefs about Donald Trump's local popularity increased their willingness to publicly express xenophobic views. Using dictator games, we also find evidence

consistent with the prediction that individuals are sanctioned less for expressing a view that is popular in their environment.

Our findings shed light on the factors that can trigger a rapid change in social norms, and in particular, norms against the expression of xenophobic views. Our results suggest that social norms regarding the expression of such views in the United States might have been causally changed by Trump's rise in popularity and eventual electoral victory.[27] More broadly, the mechanisms we study in this paper might help explain the rise, and potential consequences, of other crucial recent events such as the Brexit vote in the United Kingdom, and more generally the rise in anti-immigrant and anti-minority sentiment in the developed world.

Our analysis suggests at least two lines for future work. One deals with the joint evolution of individual views and social norms. While we see no evidence that Donald Trump's election changed people's views on immigration in the short run, it is well possible that the changed social norm will expose people to views that will eventually influence their own. These individual views could eventually affect both social norms and political decisions. Thus, understanding how individuals acquire and change their preferences through social interactions is of utmost importance. An interesting and important question, for example, is whether laws prohibiting certain speech (such as those banning denial of the Holocaust in Germany and some other countries) are more effective in forming public opinion as compared to cases where such speech is not banned but highly stigmatized (as, e.g., in the United States).

A different set of questions stems from our dictator game experiments. We observed that subjects were largely willing to forgive the individuals who publicly expressed xenophobic views in a setting where these views were expected to be more popular. Yet, they were remarkably unwilling to forgive the individual for privately expressing such views, despite knowing little about the reasons why he acquired them. This alone would be consistent with subjects viewing people from other settings as similar to them as individuals, but living in different social environments, but this explanation is perhaps too simplistic. Nevertheless, understanding how people judge thoughts and actions of people different from their own and from different societies and cultures, and perhaps ultimately why social norms emerge, is another interesting avenue for future research.

REFERENCES

Acemoglu, Daron, Tarek A. Hassan, and Ahmed Tahoun. 2018. "The Power of the Street: Evidence from Egypt's Arab Spring." *Review of Financial Studies* 31 (1): 1–42.

Acemoglu, Daron, and Matthew O. Jackson. 2017. "Social Norms and the Enforcement of Laws." *Journal of the European Economic Association* 15 (2): 245–95.

Alesina, Alberto, Paola Giuliano, and Nathan Nunn. 2013. "On the Origins of Gender Roles: Women and the Plough." *Quarterly Journal of Economics* 128 (2): 469–530.

Algan, Yann, and Pierre Cahuc. 2010. "Inherited Trust and Growth." *American Economic Review* 100 (5): 2060–92.

[27] Of course, any extrapolation of our experimental results (in which we were able to exogenously vary beliefs about the local results of the election) into a non-experimental context where elections occur naturally should be interpreted with caution. Nevertheless, the results of the motivating experiment run around the time of the 2016 election and presented in online Appendix Section D are consistent with our interpretation: while weeks right before the election there was a wedge between private and public xenophobic expression, the wedge disappeared in the week immediately following the election.

Ali, S. Nageeb, and Roland Bénabou. 2020. "Image versus Information: Changing Societal Norms and Optimal Privacy." *American Economic Journal: Microeconomics* 12 (3): 116–64.

Ali, S. Nageeb, and Charles Lin. 2013. "Why People Vote: Ethical Motives and Social Incentives." *American Economic Journal: Microeconomics* 5 (2): 73–98.

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2): 211–36.

Andreoni, James, and B. Douglas Bernheim. 2009. "Social Image and the Fifty-Fifty Norm: A Theoretical and Experimental Analysis of Audience Effects." *Econometrica* 77 (5): 1607–36.

Andreoni, James, Nikos Nikiforakis, and Simon Siegenthaler. 2017. "Social Change and the Conformity Trap." Unpublished.

Andreoni, James, Justin M. Rao, and Hannah Trachtman. 2017. "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving." *Journal of Political Economy* 125 (3): 625–53.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?" *Quarterly Journal of Economics* 124 (4): 1497-540.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78.

Bénabou, Roland, and Jean Tirole. 2011. "Laws and Norms." NBER Working Paper 17579.

Bernheim, B. Douglas. 1994. "A Theory of Conformity." *Journal of Political Economy* 102 (5): 841–77.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "Replication Data for: From Extreme to Mainstream: The Erosion of Social Norms." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E119846V1.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." Unpublished.

Bursztyn, Leonardo, and Robert Jensen. 2015. "How Does Peer Pressure Affect Educational Investments?" *Quarterly Journal of Economics* 130 (3): 1329–67.

Cho, In-Koo, and David M. Kreps. 1987. "Signaling Games and Stable Equilibria." *Quarterly Journal of Economics* 102 (2): 179–221.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50 (6): 1431–51.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127 (1): 1–56.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao. 2017. "Voting to Tell Others." *Review of Economic Studies* 84 (1): 143–81.

Elías, Julio J., Nicola Lacetera, and Mario Macis. 2019. "Paying for Kidneys? A Randomized Survey and Choice Experiment." *American Economic Review* 109 (8): 2855–88.

Enikolopov, Ruben, Alexey Makarin, Maria Petrova, and Leonid Polishchuk. 2017. "Social Image, Networks, and Protest Participation." Unpublished.

Enke, Benjamin. Forthcoming. "Moral Values and Voting." *Journal of Political Economy*.

Fernández, Raquel. 2007. "Alfred Marshall Lecture: Women, Work, and Culture." *Journal of the European Economic Association* 5 (2–3): 305–32.

Fowler, James H., and Cindy D. Kam. 2007. "Beyond the Self: Social Identity, Altruism, and Political Participation." *Journal of Politics* 69 (3): 813–27.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102 (1): 33–48.

Giani, Marco, and Pierre-Guillaume Méon. 2019. "Global Racist Contagion Following Donald Trump's Election." *British Journal of Political Science*: 1–8.

Giuliano, Paola. 2007. "Living Arrangements in Western Europe: Does Cultural Origin Matter?" *Journal of the European Economic Association* 5 (5): 927–52.

Greif, Avner. 2006. "History Lessons: The Birth of Impersonal Exchange: The Community Responsibility System and Impartial Justice." *Journal of Economic Perspectives* 20 (2): 221–36.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith. 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review* 86 (3): 653–60.

Janus, Alexander L. 2010. "The Influence of Social Desirability Pressures on Expressed Immigration Attitudes." *Social Science Quarterly* 91 (4): 928–46.

Jensen, Robert, and Emily Oster. 2009. "The Power of TV: Cable Television and Women's Status in India." *Quarterly Journal of Economics* 124 (3): 1057-94

Katz, Daniel, Floyd H. Allport, and Margaret Babcock Jenness. 1931. *Students' Attitudes: A Report of the Syracuse University Reaction Study.* Syracuse, NY: Craftsman Press.

Krupka, Erin L., and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association* 11 (3): 495–524.

Kuran, Timur. 1991. "The East European Revolution of 1989: Is It Surprising That We Were Surprised?" *American Economic Review* 81 (2): 121–25.

Kuran, Timur. 1995. *Private Truths, Public Lies: The Social Consequences of Preference Falsification.* Cambridge, MA: Harvard University Press.

Lohmann, Susanne. 1993. "A Signaling Model of Informative and Manipulative Political Action." *American Political Science Review* 87 (2): 319–33.

Lowes, Sara, Nathan Nunn, James A. Robinson, and Jonathan L. Weigel. 2017. "The Evolution of Culture and Institutions: Evidence from the Kuba Kingdom." *Econometrica* 85 (4): 1065–91.

Madestam, Andreas, Daniel Shoag, Stan Veuger, and David Yanagizawa-Drott. 2013. "Do Political Protests Matter? Evidence from the Tea Party Movement." *Quarterly Journal of Economics* 128 (4): 1633–85.

Morris, Stephen. 2001. "Political Correctness." *Journal of Political Economy* 109 (2): 231–65.

Müller, Karsten, and Carlo Schwarz. 2019. "From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment." Unpublished.

O'Gorman, Hubert J. 1975. "Pluralistic Ignorance and White Estimates of White Support for Racial Segregation." *Public Opinion Quarterly* 39 (3): 313–30.

Perez-Truglia, Ricardo, and Guillermo Cruces. 2017. "Partisan Interactions: Evidence from a Field Experiment in the United States." *Journal of Political Economy* 125 (4): 1208–43.

Prendergast, Canice. 1993. "A Theory of 'Yes Men.'" *American Economic Review* 83 (4): 757–70.

Richman, Barak D. 2017. "An Autopsy of Cooperation: Diamond Dealers and the Limits of Trust-Based Exchange." *Journal of Legal Analysis* 9 (2): 247–83.

Voigtländer, Nico, and Hans-Joachim Voth. 2012. "Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany." *Quarterly Journal of Economics* 127 (3): 1339–92.

Xiong, Heyu. Forthcoming. "Media Personality and Its Political Premium." *American Economic Journal: Applied Economics.*

This article has been cited by:

1. Tom Lane, Daniele Nosenzo, Silvia Sonderegger. 2023. Law and Norms: Empirical Evidence. *American Economic Review* **113**:5, 1255-1293. [Abstract] [View PDF article] [PDF with links]
2. Milena Djourelova. 2023. Persuasion through Slanted Language: Evidence from the Media Coverage of Immigration. *American Economic Review* **113**:3, 800-835. [Abstract] [View PDF article] [PDF with links]
3. Ingar Haaland, Christopher Roth, Johannes Wohlfart. 2023. Designing Information Provision Experiments. *Journal of Economic Literature* **61**:1, 3-40. [Abstract] [View PDF article] [PDF with links]
4. Sergei Guriev, Elias Papaioannou. 2022. The Political Economy of Populism. *Journal of Economic Literature* **60**:3, 753-832. [Abstract] [View PDF article] [PDF with links]
5. Leonardo Bursztyn, Georgy Egorov, Ingar Haaland, Aakaash Rao, Christopher Roth. 2022. Scapegoating during Crises. *AEA Papers and Proceedings* **112**, 151-155. [Abstract] [View PDF article] [PDF with links]