I Have Nothing Against Them, But...

Leonardo Bursztyn
Ingar Haaland
Aakaash Rao
Christopher Roth*

March 2, 2020

Abstract

We study the role of commonly known justifications in facilitating the public expression of stigmatized views, such as opposition to a minority group. In our theoretical framework, agents may take an anti-minority action for two reasons: because they are intolerant, or because they are gullible and easily persuaded by misinformation. The availability of a commonly known justification to oppose minorities allows intolerant, non-gullible agents to take the anti-minority action while pooling with tolerant, gullible agents to avoid social stigma. We test our model's predictions through two related experiments. In the first experiment, we show that participants believe that donors who had been exposed to a rationale for publicly donating to an anti-immigrant organization are less biased against foreigners and more gullible than donors who had not been exposed. In the second experiment, we vary the availability of a commonly known justification for donating to the anti-immigrant organization, and we show that the availability of the justification significantly increases participants' willingness to donate. We conclude by discussing implications of our framework for debunking fake news on social media and understanding populist rhetoric.

*Bursztyn: University of Chicago and NBER. Email: bursztyn@uchicago.edu. Haaland: University of Bergen and CESifo. Email: ingar.haaland@uib.no. Rao: Harvard University. Email: arao@g.harvard.edu. Roth: University of Warwick, briq, CEPR, CESifo, and CAGE. Email: christopher.roth@warwick.ac.uk. We thank Daron Acemoglu, Alberto Alesina, Ben Enke, Ed Glaeser, Emir Kamenica, Ross Matheis, Nathan Nunn, and Andrei Shleifer for very helpful suggestions. We thank Takuma Habu and Alison Zhao for outstanding research assistance. We gratefully acknowledge financial support from the Pearson Institute for the Study and Resolution of Global Conflicts and the University of Chicago Social Sciences Research Center. The experiments were pre-registered on the AEA RCT registry with IDs AEARCTR-0005308 and AEARCTR-0005462. The research described in this article was approved by the University of Chicago Social and Behavioral Sciences Institutional Review Board (protocol IRB19-1320) and the Humanities and Social Sciences Research Ethics Committee at the University of Warwick (protocol HSSREC 76/19-20). Replication files are publicly available at www.github.com/AakaashRao/excuses.

1 Introduction

How do anti-minority movements gain traction in settings where public expression of resentment towards minorities is stigmatized? Propaganda, political rhetoric, and mass media appear to have substantial effects on people's willingness to engage in xenophobic actions against immigrants and other minorities (Enikolopov and Petrova, 2015; Zhuravskaya et al., 2019). For example, radio propaganda caused increased killings during the Rwandan genocide (Yanagizawa-Drott, 2014) and anti-Semitic acts in Nazi Germany (Adena et al., 2015), while inflammatory Tweets from Donald Trump cause increased anti-Muslim hate crimes (Müller and Schwarz, 2018).

The most commonly proposed mechanism is that mass media affects behavior through persuasion: people are more willing to publicly take anti-minority actions because the information disseminated through media has changed their private opinions. However, this explanation is difficult to reconcile with a growing body of survey experimental evidence suggesting mostly small or null effects of information on stated preferences and behavior (Alesina et al., 2018; Grigorieff et al., 2020; Hopkins et al., 2019; Kuziemko et al., 2015). While a number of factors differ between field and lab settings, making direct comparison impossible, the discrepancy suggests that media may influence behavior through mechanisms other than persuasion.

In this paper, we propose an alternative mechanism through which mass media can affect public behavior. Commonly known justifications, whether true or false, can be used as an "excuse" to engage in otherwise-stigmatized behavior insofar as they lower the associated social stigma. For example, consider people who oppose immigration from Mexico simply because they dislike Mexicans, yet cannot express this rationale in most public settings without incurring social costs. Once an anti-Mexican justification (e.g. that Mexican immigrants bring crime into the country) becomes common knowledge — perhaps because a political candidate bases a campaign around anti-Mexican rhetoric — these people are given an "excuse": they can attribute their opposition to Mexican immigration to their believing the anti-Mexican narrative, even if they privately know it to be false and *even if they know others know it to be false*. As long as being seen as gullible is less socially costly than being seen as intolerant, the common knowledge of plausible justifications can increase the equilibrium level of public anti-minority expression.

We formalize this intuition with a simple theoretical model. Agents differ on two dimensions: tolerance (intolerant vs. tolerant) and sophistication (gullible vs. sophisticated). The true state of the world is either that immigration is harmful or that immigration is not harmful, and an anti-immigrant story (which may be true or false, depending on the true state of the world) is spread throughout the society. Agents receive a signal about the veracity of the story; sophisticated agents' signals are perfectly informative about the true state of the world, while gullible agents' signals are noisy: they may receive a signal that the story is true even if it is not.[1] Each agent must

[1] All agents in our model are Bayesian, and thus gullible agents understand that their signals may be incorrect and adjust their inferences accordingly.

then decide whether to take an anti-immigrant action. A Bayesian audience observes the agent's action and uses this information to make an inference about the agent's type conditional upon their action.

We distinguish between two conditions, which map cleanly onto our two experiments. In the "Excuse" condition, the audience correctly believes that the agent received a signal, and thus that he may have donated either because he is intolerant or because he is gullible and received an anti-immigrant signal. In contrast, in the "No Excuse" condition, the audience incorrectly believes that the agent did *not* receive a signal, and thus that any differences in donation behavior are due entirely to differences in tolerance. Agents receive both expressive utility from acting in a manner consistent with their political beliefs and reputational utility from acting in a socially rewarded manner. In the state of the world where immigration is not harmful, because tolerant and gullible agents may receive an anti-immigrant signal and rationally choose to donate, intolerant and sophisticated agents — who correctly perceive that immigration is not harmful — can pool with the tolerant and gullible agents by donating. In essence, because being gullible is less stigmatized than being intolerant, intolerant and sophisticated agents can pretend that they have been persuaded by their signal in order to avoid the stigma associated with being intolerant.

We test the key predictions of our model using two complementary large-scale experiments. In Experiment 1, conducted with a broadly representative sample of 5065 Democrats, we study whether participants hold different beliefs about a matched partner's bias and about that partner's gullibility when that partner had a justification for taking an anti-immigrant action, relative to when they did not have a justification. We inform our respondents that we have matched them with a respondent from another study who took an anti-immigrant action (authorizing a donation to "Fund the Wall"). All respondents in this experiment are informed about a recent study (Lott, 2018), which finds that undocumented immigrants in Arizona commit crimes at substantially higher rates than comparable US citizens.[2] We also inform them that the Cato Institute, a nonpartisan libertarian think-tank, has challenged the validity of the study.[3] Our key experimental variation is whether our respondents believe that their matched respondent knew about the study before making their decision: while respondents in the "Excuse" condition are told that their matched respondent knew about the study before making their decision, respondents in the "No Excuse" condition are told that their matched respondent was not told about the study.

To most closely capture the natural process of inference and to avoid priming respondents,

[2]The Trump administration has cited this study repeatedly as evidence for the impact of illegal immigration on crime. For example, in a January 2018 speech on "national security and immigration priorities of the administration," then-Attorney General Jeff Sessions claimed that the study proved that "tens of thousands of crimes have been committed in this country that would never have happened if our immigration laws were enforced and respected like they ought to be" (Sessions, 2018).

[3]Additionally, to further ensure that they are not left with a distorted view of the relationship between immigration and crime, we debrief respondents at the end of the survey by providing them with a short summary of the empirical evidence on the effects of immigration on crime and a link to a relevant meta-analysis. In contrast to the study we cite, most work generally find that undocumented immigrants commit crimes at rates similar to or lower than comparable US citizens.

we first measure participants' beliefs about their matched respondents' motives for donating to Fund the Wall using an open-ended question, directly measuring what "comes to mind". We first use text analysis to measure the extent to which participants ascribe their matched respondents' motives to gullibility as compared to intolerance. We then turn to a more structured measure of beliefs: half of the participants make an incentivized guess about their matched respondent's score on a test measuring gullibility; the other half make an incentivized guess about their matched respondent's score on a test measuring cultural tolerance.

Consistent with our model's predictions, we find strong treatment effects on both measures. In describing why they believed their matched respondent chose to donate to Fund the Wall, participants matched with a respondent who had no justification for donating are 7 percentage points (70%) more likely to use a word related to intolerance (p <0.001) and 3 percentage points (43%) less likely to use a word related to gullibility ($p < 0.001$) relative to participants matched with a respondent who had a justification. We find similar treatment effects on the structured belief measures: participants believe that a matched respondent with an excuse scored 0.14 standard deviations lower on the intolerance scale ($p < 0.001$), and 0.32 standard deviations higher score on the gullibility scale ($p < 0.001$). Taken together, our evidence from Experiment 1 suggests that publicly known justifications for xenophobic behavior can lower the social image costs associated with such behavior.

This finding begs the question of whether people who are intolerant types strategically use commonly known justifications to avoid the social stigma associated with publicly expressing their views. We investigate this question in Experiment 2, in which we recruit a broadly representative sample of 4,432 Republicans and Independents and study whether they are more willing to publicly authorize a donation to Fund the Wall when they have a justification for doing so.

Our experiment proceeds as follows. We begin by informing participants about the Lott study, and we give them the opportunity to authorize a $1 donation to Fund the Wall. We tell participants that we will post their individual donation decisions on our website, and that in order to communicate our research findings to the public, we will publicize the website among residents in their city. This generates a real social cost of authorizing a donation, particularly in areas where respondents expect the majority of the population to disapprove of the donation.

Identifying the "excuse effect" of justifications requires disentangling it both from the direct effect of persuasion (first-order beliefs) and from a change in anticipated social approval associated with the potential for the justification to change the views of others (second-order beliefs). We hold first-and second-order beliefs fixed across the "Excuse" and "No Excuse" condition by (1) informing participants in both conditions about the study and (2) by making it clear that the website on which their donation decisions will be posted will also contain information about the study, such that all visitors will learn about the study before viewing individual donation decisions. The key treatment varies the availability of a justification for donating. In particular, participants in the Excuse condition see that their audience will learn that they knew about the Lott study when

making the donation decision, while participants in the No Excuse condition believe that their audience will *not* know that they knew about the Lott study.

We find a large and statistically significant excuse effect on the participants' willingness to publicly donate to Fund the Wall. Participants in the Excuse condition are 6.3 percentage points (13%) more likely to authorize the donation than respondents in the No Excuse condition ($p < 0.001$). To benchmark the effect size, we compare the donation rate in a control condition — in which participants are not informed about the study, and implicitly believe that website visitors will also not be informed — with the donation rate in the No Excuse condition, which allows us to identify the joint effect of first- and second- order beliefs. We find that this joint effect is small relative to the effect of third-order beliefs, highlighting a quantitatively important role of excuses relative to the effect of persuasion and anticipated persuasion of one's audience. Moreover, the effect is driven by participants who live in majority-Democrat areas, consistent with the prediction that participants will more strongly require a justification for their action when their audience is likely to disapprove of it.

Our results contribute to a growing empirical and theoretical literature on the effects of image concerns on economic and moral decision-making (Bénabou and Tirole, 2006; Bénabou et al., 2018). Prior work has studied the effect of image concerns on political and economic outcomes (including voting, as in DellaVigna et al. 2017; campaign donations, as in Perez-Truglia and Cruces 2017; educational investments, as in Bursztyn et al. 2019a and Bursztyn et al. 2017; health investments, as in Karing 2018; and political activism, as in Cantoni et al. 2019 and Hager et al. 2019). Kuran 1997 argues that "preference falsification" — expressing a public view distinct from (often opposite of) one's private view in order to conform to perceived social norms — can have dramatic consequences for political equilibria. Bursztyn et al. (2019b) show that updating views about Donald Trump's popularity, either experimentally or through the "natural experiment" of Trump's electoral victory, eliminated the gap between public and private support for an anti-immigrant organization. Similarly, Bursztyn et al. (2018) find that experimentally correcting misperceptions about the acceptability of female labor force participation increases the probability that a woman will accept a job outside of the home. Relative to existing work, which generally highlights a single type dimension on which respondents signal and update (e.g. civic-mindedness, as in DellaVigna et al. 2017, or pro-sociality, as in Ariely et al. 2009), a key contribution of this paper is to show that people can strategically use information to influence which of two dimensions on which an audience updates. We directly test our model's theoretical prediction employing a novel approach to measure inference about types, using both text analysis and structured belief measures. Moreover, while the previous literature on image concerns more broadly focuses on the importance of second-order beliefs, our evidence on xenophobic expression provides novel evidence on the importance of third-order beliefs (beliefs about others' beliefs about one's own beliefs). We thus emphasize the importance of commonly known justifications, which can be created by political entrepreneurs and the media, in facilitating xenophobic expression.

Our work also adds to a literature on the effects of media and propaganda on political and economic behavior (Banerjee et al., 2019; Bursztyn et al., 2020; DellaVigna and La Ferrara, 2015; La Ferrara, 2016; La Ferrara et al., 2012). Work on the effects of media on violence (Müller and Schwarz, 2018; Yanagizawa-Drott, 2014), protest participation (Enikolopov et al., 2019), and other public outcomes are generally unable to distinguish between direct persuasion and social image concerns arising from changes in higher-order beliefs. We demonstrate that media can generate justifications to violate social norms, leading to changes in public behavior even in the absence of persuasive impact.

Finally, our work relates to a literature on moral "wiggle room". Robust laboratory evidence (e.g. Dana et al. 2007; Hamman et al. 2010; Lazear et al. 2012) shows that the availability of even weak justifications to behave selfishly (e.g. choosing not to click a button to reveal a partner's payoffs) has substantial effects on behavior. Exley (2016) finds that individuals use risk as a justification to avoid donating to charitable causes. Because decisions in these settings are anonymous, these findings can be rationalized by a behavioral model of self-signaling, as in Bénabou and Tirole (2011) or by social image concerns vis a vis the experimenter. Other work has studied settings where decisions are observable, generating image concerns. Andreoni and Bernheim (2009) find that increasing the probability that the dictator's choice will be ignored and the recipient allocated an unfavorable amount reduces generosity by giving the dictator "plausible deniability" vis-a-vis the recipient. Conversely, Ariely et al. (2009) show that extrinsic incentives for prosocial behavior can crowd out image motivation (a possibility suggested by Bénabou and Tirole (2006)), which can in our framework be interpreted as a "reverse excuse" that decreases the extent to which the audience updates about an agent's prosociality in light of a prosocial action. Our work contributes to this literature by formalizing a two-type equilibrium signaling model, highlighting the use of commonly known excuses to generate moral wiggle room in a relevant natural setting.

The remainder of this paper proceeds as follows. In Section 2, we outline a model of communication in which the availability of excuses allows xenophobic, sophisticated agents to pool with tolerant, gullible agents when taking a publicly stigmatized action. In Section 3, we present an online experiment testing our model's first prediction as to how the interpretation of xenophobic actions depends on the availability of excuses. In Section 4, we present a second online experiment testing the model's second prediction, quantifying the effect of third-order beliefs on public expression and benchmarking its magnitude against the combined effects of first- and second-order beliefs. Section 5 discusses robustness of our experimental findings to attrition and experimenter demand, presenting a novel machine learning approach to measuring experimenter demand effects. Section 6 discusses implications of our framework for understanding populist rhetoric, and the spread of fake news on social media. Section 7 concludes.

2 Theoretical Framework

A society consists of a continuum of agents. Agents differ on two dimensions: their tolerance toward foreign cultures (either High or Low, i.e. $\theta \in \{H, L\}$) and their sophistication (either Sophisticated or Gullible, with $\gamma_i \in \{S, N\}$). Thus, agents are characterized by their type: $\tau_i = \{\theta_i, \gamma_i\}$, which is privately known. The distributions from which both dimensions are drawn are independent and common knowledge, with $\Pr(\theta_i = H) = \pi > 1/2$ and $\Pr(\gamma_i = S) = \delta$.

At the beginning of the game, an anti-immigrant news story is spread throughout the society. With probability p, the story is true ($n = T$), while with probability $1 - p$, the story is false ($n = F$). p is common knowledge. All agents then receive a signal $s \in \{T, F\}$ about the veracity of the story. Sophisticated agents' signals are perfectly informative about the veracity of the story: $\Pr(s_i^S = T|n = T) = \Pr(s_i^S = F|n = F) = 1$. However, gullible agents receive a biased signal of the true state of the world. While $\Pr(s_i^N = T|n = T) = 1$, so that gullible agents always receive a true signal when the story is true, $\Pr(s_i^N = T|n = F) = q$, so that gullible agents may receive a true signal even when the story is false.

After observing their signals, each agent simultaneously chooses whether or not to take an anti-immigration action: $a \in \{0, 1\}$.

Agents receive private utility from donating, which has two components: one component R_i independent of their beliefs about the true state of the world ("acting according to one's principles"), and another component D_i dependent on their beliefs about the true state of the world ("taking the correct action"). $R_i = R_H$ for $\theta_i = H$ and $R_i = R_L$ for $\theta_i = L$, and we assume that $R_H < c < R_L$, i.e. tolerant agents receive lower utility than intolerant agents from donating, regardless of their beliefs about the veracity of the news. $D_i = [a \Pr(n = F|s_i) + (1 - a) \Pr(n = T|s_i)]c$: agents face no cost of donating when they believe donating is the "correct action" (that is, when they are certain that the story is true) and they face the full cost c from donating when they are certain that donating is the "incorrect action" (that is, when they are certain that the story is false). Similarly, they face the full cost of not donating when they believe donating is the "correct action"; and they face no cost of not donating when they believe donating is the "incorrect action".

Agents also receive social utility ∇_i based on how they are judged by others. Each agent i (the "sender") is randomly matched with another agent j (the "receiver"), who is drawn from the same distribution of types and who also receives a signal about the true state of the world. The receiver rewards the sender based on his perceived probability that the sender is of the same tolerance type: $\nabla_i = \lambda \Pr_j(\theta_i = \theta_j|Y)$, where Y refers to the information set available to the receiver (which varies across experimental conditions). We assume that $\lambda > \frac{R_L + c}{(2\pi - 1)\pi}$.

Thus, the sender's utility function is as follows:

$$u_i = a[R_i - c\Pr_i(n = F|s_i)] - (1 - a)[c\Pr_i(n = T|s_i)] + \lambda E_i(\Pr_j(\theta_i = \theta_j|Y)) \quad (1)$$

To map to our experiment, we consider two conditions: the "no excuse" condition (NE) and "excuse" condition (E). In the "no excuse" condition, the receiver observes the sender's action and observes his or her own signal. The receiver believes that senders do not receive any signal. Thus, $Y_{NE} = \{a_i, s_j\}$. In the "excuse" condition, the receiver additionally believes that his or her matched sender has a signal s_i following its regular distribution Ω. Thus, $Y_E = \{a_i, \Omega(s_i), s_j\}$.

Following the set-up, sophisticated agents know that their signals s_i are perfectly informative of the veracity of the story, while the gullible agents rationally account for the potential that they received an incorrect signal: $\Pr(n = T|s^N = T) = \frac{p}{p+q(1-p)}$ and $\Pr(n = T|s^N = F) = 0$. We denote the probability of a sender of morality type θ_i and sophistication type γ_i with signal s_i choosing $a = 1$ by $\rho^{s_i}_{\theta_i}(\gamma_i)$; for example, the probability of a high-morality, sophisticated agent choosing $a = 1$ with signal $s_i = T$ is written as $\rho^T_H(S)$. We adopt the Perfect Bayesian Equilibrium as the solution concept for the game, and we further adopt the D1 criteria (Cho and Sobel, 1990) to characterize the out-of-path equilibria.

2.1 "No Excuse" condition

Since the receiver believes that senders do not observe a signal, she updates her inference about the sender's type solely based on the sender's action. Therefore, regardless of the sender's sophistication level and the signal he truly receives, his social reward term is denoted by $\lambda\nabla^a_i$. In Appendix B.1.2, we show that the following inequality holds:

$$R_i + c + \lambda(\widetilde{\nabla}^1_i - \widetilde{\nabla}^0_i) > R_i + \frac{p - q(1-p)}{p + q(1-p)}c + \lambda(\widetilde{\nabla}^1_i - \widetilde{\nabla}^0_i). \tag{2}$$

That is, holding fixed the sender's tolerance type, a sophisticated sender receiving signal s_i is always weakly more likely to choose $a = 1$ than a gullible sender receiving the same signal.

We also have $u(a = 1) - u(a = 0) = R_i - c + \lambda(\widetilde{\nabla}^1_i - \widetilde{\nabla}^0_i)$ for both sophisticated and gullible senders when they receive a false signal.

As we show in Appendix B.1.2:

$$R_i + \frac{p - q(1-p)}{p + q(1-p)}c + \lambda(\widetilde{\nabla}^1_i - \widetilde{\nabla}^0_i) > R_i - c + \lambda(\widetilde{\nabla}^1_i - \widetilde{\nabla}^0_i), \tag{3}$$

that is, a sender receiving a true signal is always more likely to choose $a = 1$ than a sender receiving a false signal.

Thus, $\rho^{s_i}_L(\gamma_i) \geq \rho^{s_i}_H(\gamma_i)$: fixing the sender's sophistication level and the signal he receives, he is more likely to choose $a = 1$ when he is of the intolerant type since $R_H < R_L$.

Since $R_i + c + \lambda(\widetilde{\nabla}^1_i - \widetilde{\nabla}^0_i) < 0$ under the assumption $\lambda > \frac{R_L + c}{(2\pi - 1)\pi}$, i.e. $\rho^T_L(S) = 0$, donating is strictly dominated for all types: the loss sender i suffers from donating always outweighs the expressive utility he obtains from donating, regardless of his tolerance type, sophistication level,

and the signal he receives.

2.2 "Excuse" condition

In the Excuse condition, it is common knowledge that senders have received a signal. Thus, both the sender and the receiver update their beliefs about the signals they have received. Letting i denote the sender and j denote the receiver, and fixing the notation $\mathbb{M}(a = x) = \Pr_j(\theta = H|a = x)$, we can write the sender's utility function as follows:

$$
\begin{aligned}
u_i =& a[R_i - cPr_i(n = F|s_i)] - (1 - a)[cPr_i(n = T|s_i)] + \\
& \lambda \Bigg[\pi \Big(Pr_i(s_j = T|s_i)(\delta \mathbb{M}_T^S(a = 1) + (1 - \delta)\mathbb{M}_T^N(a = 1)) + Pr_i(s_j = F|s_i)\mathbb{M}_F(a = 1) \Big) + \\
& (1 - \pi) \Big(Pr_i(s_j = T|s_i)(\delta(1 - \mathbb{M}_T^S(a = 1)) + (1 - \delta)(1 - \mathbb{M}_T^N(a = 1))) + Pr_i(s_j = F|s_i)(1 - \mathbb{M}_F(a = 1)) \Big) \Bigg]
\end{aligned} \tag{4}
$$

As we formalize in Appendix B.2.1, the common knowledge that the sender has received a signal changes the equilibrium inference about the sender's type. In particular, the following proposition holds:

Proposition 1 *Receivers believe that senders who took the anti-immigrant action in the No Excuse condition are on average less tolerant and more gullible than senders who took the anti-immigrant action in the Excuse condition. That is,*

$$\Pr_j[\theta_i = H|a_i = 1, s_j] \leq \Pr_j[\theta_i = H|a_i = 1, \Omega(s_i), s_j], \tag{5}$$

and

$$\Pr_j[\gamma_i = G|a_i = 1, s_j] \geq \Pr_j[\gamma_i = G|a_i = 1, \Omega(s_i), s_j]. \tag{6}$$

The inequality is strict for $\lambda \in \left(\frac{R_L + c}{(2\pi - 1)\pi}, \frac{R_L + c}{\pi} \right)$.

As a result, senders are more willing to donate, since the associated social image cost is lower. As we show in Appendix B.2.2, the following proposition holds:

Proposition 2 *The share of senders taking the anti-immigrant action in the No Excuse condition is weakly lower than that in the Excuse condition. That is,*

$$\Pr[a_i = 1|Y_j = (a_i, s_j)] \leq \Pr[a_i = 1|Y_j = (a_i, \Omega(s_i), s_j)]. \tag{7}$$

The inequality is strict for $\lambda \in \left(\frac{R_L+c}{(2\pi-1)\pi}, \frac{R_L+c}{\pi}\right)$.

The intuition for the strict inequality in Proposition 2 is straightforward. When the social reward parameter λ is too low, agents are insensitive to the presence of the excuse because their private reward (which is held fixed across the No Excuse and Excuse condition) dominates their social reward. When the social reward parameter λ is too high, agents are again insensitive to the presence of the excuse because all agents take the more socially acceptable action (i.e. they do not take the anti-immigrant action) whether or not they have the excuse.

3 Excuses and Type Inference

We first test our model's predictions about type inference (Proposition 1) by studying how the availability of a justification changes how an audience interprets the decision to donate to Fund the Wall.[4] We are particularly interested in how excuses for an action affect judgment vis-a-vis an audience who *disapproves* of the action, as this is precisely the audience before which an agent may require an excuse. We thus focus on Democrats, who are most likely to disapprove of the decision to donate to Fund the Wall.[5] As we show in Experiment 2, public behavior among people who live in counties with a lower Republican vote share is substantially more elastic to the availability of an excuse than public behavior among those who live in more Republican counties, suggesting that non-Republicans are indeed the relevant audience to consider when studying how excuses affect inference.

3.1 Sample

We conducted a pre-registered experiment with Luc.id, a widely used online panel provider (Wood and Porter, 2019). We recruited a sample of 5,065 Democrats in February 2020.[6] Participants were directed to our survey on the online platform *Qualtrics*. Only participants who were over the age of 18, resided in the United States, indicated their consent to participate, and passed a simple test of attention were allowed to proceed. Our sample of respondents is broadly representative of Democrats in the United States (Appendix Table A) and well-balanced on observables across treatment arms (Appendix Table A2).

3.2 Experimental design

We tell all respondents about a recent study (Lott, 2018) which finds that that "undocumented immigrants are at least 142% more likely to be convicted of a crime than other Arizonans. They

[4] All survey instruments are available in Appendix C.

[5] As of January 2019, 6 percent of Democrats or Democratic leaners favored "substantially expanding the wall", compared to 82 percent of Republicans or Republican leaners (Pew, 2019).

[6] In our pre-registration, we specified that in some specifications, we would pool data from a pilot with the pre-registered data. We report both unpooled and pooled specifications.

also tend to commit more serious crimes and serve 10.5% longer sentences, are more likely to be classified as dangerous, and 45% more likely to be gang members than U.S. citizens."[7] We also truthfully tell our respondents that a number of sources (including a researcher affiliated with the Cato Institute, a non-partisan think tank) have recently challenged some of the study's methods, claiming that errors in analysis invalidate its results.[11]

We then truthfully tell participants that we conducted a project on political and social attitudes in the United States earlier in the year, and that respondents to this previous study were given an opportunity to authorize a $1 donation to Fund the Wall, a nonprofit organization that seeks to reduce illegal immigration into the United States by helping to fund and construct the US-Mexico border wall. We make it clear that the respondents from this survey knew that their donation decision would be posted on our study website. We inform participants that we have matched them with one of these respondents, and that this respondent chose to authorize the donation. Respondents in the Excuse condition are truthfully told that their matched respondent was informed about the study before deciding whether or not to authorize the donation to Fund the Wall, while respondents in the No Excuse condition learn that their matched respondent was not informed about the study before making their donation decision.

Measuring type inference After learning whether or not their partner learned about the study, all participants respond to the following open-ended question: "Why do you think your matched respondent chose to donate to Fund the Wall?" As we will discuss in Section 3.3, these open-ended responses form the raw data for our first measure of type inference; we employ text analysis to systematically analyze the open-ended responses. Participants are then cross-randomized into one of two conditions: "tolerance" and "gullbility". Participants in the "tolerance" condition are truthfully told that their matched respondent completed the "Foreign Culture Tolerance Scale," a "short questionnaire measuring tolerance toward foreign values and traditions". Participants in the "gullibility" condition are truthfully told that their matched respondent completed the "Gullibility Scale," a "short questionnaire which measures how easily people are manipulated by evidence from untrustworthy sources". All participants are asked to guess their respondent's score; we incentivize this guess by informing them that if they correctly guess the score, they will be entered into a lottery for a $50 Amazon gift card.

[7] This study has been widely covered by the media, including *The Washington Times*[8], *National Review*[9], and *Fox News*[10] and has been repeatedly cited by Trump administration officials. For example, in a January 2018 speech on "national security and immigration priorities of the administration," then-Attorney General Jeff Sessions claimed that the study proved that "tens of thousands of crimes have been committed in this country that would never have happened if our immigration laws were enforced and respected like they ought to be" (Sessions, 2018).

[11] In order to ensure that our respondents are not misinformed, we debrief them at the end of the study and provide them with a meta-analysis summarizing the work on the effects of immigration on crime.

3.3 Main results

Empirical strategy To identify the effect of the excuse on respondents' inference about the matched respondent's type, we estimate the following empirical specification:

$$y_i = \alpha_0 + \alpha_1 \text{Excuse}_i + \varepsilon_i, \tag{8}$$

where Excuse_i is an indicator taking value 1 for participants in the Excuse condition and value 0 in the No Excuse condition. y_i is our participant's belief about the matched respondent's type. We employ robust standard errors throughout.

Main findings We begin by using text analysis to measure how participants respond to the open-ended question "Why do you think your matched respondent chose to donate to Fund the Wall?" The advantage of this approach is that we can directly measure what comes to respondents' minds rather than priming them on the particular dimensions in which we are interested. Measuring type inference through analyzing open-ended text responses may thus better capture the natural process of inference than directly asking about perceptions of tolerance or gullibility.

We began with five "seed words" for each type. For (in)tolerance, we chose *racist*, *biased*, *xenophobic*, *intolerant*, and *prejudiced*. For gullibility, we chose *convinced*, *persuaded*, *gullible*, *gullible*, and *sucker*. We added all "most relevant" synonyms for these words, as marked on the website `www.thesaurus.com`. In order to capture different parts of speech, we stem all of the words, we then stemmed all words in our list (e.g. xenophobic $\rightarrow$ xenophob, gullible $\rightarrow$ gullib), for a total of 23 intolerance-related stems and 30 gullibility-related stems (Gentzkow et al., 2017).

We then define two indicator variables — one variable that takes value 1 if the respondent uses an intolerance-related stem and 0 otherwise; and another variable that takes value 1 if the respondent uses a gullibility-related stem and 0 otherwise – and estimate treatment effects on the probability that the respondent uses at least one word in each list. In order to eliminate potential experimenter degrees of freedom, we pre-specified this entire procedure, including the list of stems and the code file used for analysis.

Table 1 and Figure 7 display the results from our text-based type inference. Participants in the Excuse condition are 7 percentage points less likely to use a stem related to intolerance when describing their matched partner's motive, compared to a mean of 17 percent among participants in the No Excuse condition ($p < 0.001$). These same participants are also 3 percentage points more likely to use words related to gullibility ($p < 0.001$), relative to a mean of 7 percent among participants in the No Excuse condition.[12] These are substantial effect sizes, which highlight that the availability of a justification strongly changes people's inference about their matched respondent's motives.

[12] We were intentionally conservative when choosing stem words in order to minimize the rate of false positives.

Table 2 and Figure 7 display the main findings of Experiment 2 based on the structured belief measures. Participants who believe that their matched respondent had an excuse had a 0.13 standard deviation lower score on the intolerance scale ($p < 0.001$), and a 0.32 standard deviation higher score on the gullibility scale ($p < 0.001$). These effects are very similar in the pilot and in the pre-registered main experiment, are robust to the inclusion of control variables, and are precisely estimated. Moreover, the text-based measures of type inference are highly correlated with the scale-based measures, as shown in Table A3. On average, a respondent who uses a word related to intolerance (gullibility) when describing the matched respondent's motive rates the matched respondent as half a standard deviation more intolerant (gullible) than a respondent who does not use such a word.

Taken together, our evidence suggests that when judging others' motives, people believe that those who donated with an excuse are more gullible and less intolerant than those who donated without an excuse. Our results thus confirm the first key prediction of our signaling model.

4 Excuses and Xenophobic Expression

Our results in Experiment 1 beg the question of whether agents strategically use commonly known justifications to disguise their bias against immigrants, i.e. Proposition 2 from our model. In this section, we present the results of a second experiment testing this hypothesis. Because pilot results suggested that the base rate at which Democrats authorized a donation to Fund the Wall was extremely low, we focused in this experiment on Republicans and Independents.

4.1 Sample

Sample and survey administration We once again worked with Luc.id to recruit 3750 self-reported Republicans and Independents in a pre-specified experiment in January 2020. We supplement this data with approximately 800 Republicans and Independents from a pilot experiment with Luc.id, also conducted in January 2020. Participants were directed to our survey on the online platform *Qualtrics*. As before, only participants who were over the age of 18, resided in the United States, indicated their consent to participate, and passed a simple test of attention were allowed to proceed.[13] Our sample of respondents is broadly representative of Independents and Republicans in the United States (Appendix Table A) and well-balanced in terms of observables across treatment arms (Appendix Table A7).

4.2 Experimental design

Information: Lott study After completing a series of demographic questions, respondents are assigned to one of three different treatment conditions: an Excuse condition, a No Excuse con-

[13] All survey instruments are available in Appendix D.

dition, and a Control. Respondents in the Excuse and No Excuse condition receive the same information as respondents from Experiment 1: they are told about the Lott study, which finds that immigrants commit more crimes and more serious crimes than US citizens, and about the fact that the study has been challenged by a variety of sources. Respondents in the Control condition do not learn about the study.

Introduction: donation decisions To minimize experimenter demand concerns, we truthfully tell our respondents that we will randomly select one of two organizations, and the respondents will have the opportunity to authorize a $1 donation to this organization. In practice, we randomized almost all of our respondents to Fund the Wall.

Visibility manipulation We ask respondents to give permission for us to access their names, city, and their operating system from the survey provider. (The survey company indeed has access to the participants' personal identifying information and was willing to share this information, but to protect participant privacy, we did not access it.) We then tell our respondents that we will post the results from this survey, including their "individual donation decision" and the donation decisions of all of the other respondents to this survey on our website.[14]

We inform our respondents that "we believe it is important to communicate our findings about political and social attitudes in [City of respondent] to the public".[15] We then inform our respondents that "we will work with major news organizations in [City of respondent] with both a liberal and conservative viewership to publicize our website through newspaper and website articles", and "we will also promote our website via Facebook ads to [City of respondent] residents". This generates a plausible social cost for acting in a way that will be stigmatized in the respondent's area.

Varying third-order beliefs Our main object of interest is to compare individuals who hold identical first- and second- order beliefs, but different third-order beliefs. To hold first-order beliefs constant, all respondents in the Excuse and No Excuse condition received information about the study. To hold second order beliefs constant, we show respondents in the Excuse and No Excuse conditions screenshots of our website, clearly indicating that all website visitors will be informed about Dr Lott's study. To non-deceptively vary third-order beliefs, we exploit the fact that Lott's study has not yet been published (i.e. it is currently a working paper on SSRN).

In particular, in the Excuse treatment, we write that "Website visitors will know that you knew about the results of Dr. Lott's study," giving respondents an excuse to donate (i.e. believing, based on the findings of the Lott study, that illegal immigrants commit substantially more crime than citizens). In the No Excuse treatment, we write that "Website visitors will believe that you did

[14]To avoid explicit deception without compromising respondents' privacy, we post individual donation decisions linked with study IDs rather than names.

[15]We used IP address to capture and display their current location (i.e. their city).

not know about Dr. Lott's study, since the website states that you were surveyed before the study was published and does not mention that you were shown a summary of the study's findings". Respondents in this condition thus believe that their audience will believe that they (respondents) had no information excusing their decision to donate to fund the border wall. As forming correct higher-order beliefs may be cognitively challenging, we included an illustration of their audience's information set tailored to each of the three different treatment conditions (Figure A1).

4.3 Main results

Empirical strategy To identify the joint effects of first- and second-order beliefs (i.e. the direct persuasive effect of learning about the Lott study in addition to the indirect effect of learning that one's audience has learned about the Lott study and is thus more likely to approve of the donation), we compare the Control condition with the No Excuse condition. To identify the effect of third-order beliefs, we compare the No Excuse condition to the Excuse condition. This design thus allows us to benchmark the effect of third-order beliefs against the combined effect of first- and second-order beliefs. Thus, our main specification of interest is given as follows:

$$y_i = \beta_0 + \beta_1 \text{Excuse}_i + \beta_2 \text{Control}_i + \varepsilon_i \tag{9}$$

where y_i is an indicator taking value 1 if the respondent authorized the donation to Fund the Wall and 0 otherwise; Excuse_i is an indicator taking value 1 if the respondent was assigned to the Excuse condition and 0 otherwise; and Control_i is an indicator taking value 1 if the respondent was assigned to the Control condition and 0 otherwise. The omitted category is thus the No Excuse condition. We employ robust standard errors throughout our analysis.

Main findings Table 3 and Figure 2 display the main findings of Experiment 1. We find a large and statistically significant effect of third-order beliefs on people's tendency to authorize a donation to Fund the Wall. Respondents in the Excuse condition are 6.3 percentage points more likely to authorize the donation compared to respondents in the No Excuse condition. This effect is highly statistically significant ($p < 0.001$), and large relative to a Control condition mean of 48.8 percentage points. These effects are very robust to the inclusion of control variables and are almost identical in our pre-specified main study and a pilot study.

To benchmark the effect sizes of the excuse effect, we are interested in the joint effect of first-order beliefs and second-order beliefs on making public donations to Fund the Wall. To identify this joint effect, we compare respondents in the Control condition with respondents in the No Excuse condition. Respondents in the No Excuse condition are approximately 0.007 percentage points more likely to donate to Fund the Wall relative to respondents in the Control condition.

This effect represents the combined effects of persuasion and anticipated persuasion of one's

audience. Relatively small persuasion effects are in line with other information provision experiments in the immigration domain, which typically find small or null effects on behavior and stated preferences (Alesina et al., 2018; Grigorieff et al., 2020; Hopkins et al., 2019). Thus, small effects of anticipated persuasion are consistent with agents holding accurate expectations about whether their audience will be persuaded. Given the small joint effect of persuasion and anticipated persuasion, what might explain the large excuse effect we observe? First, agents may simply hold incorrect higher-order beliefs: in particular, they may believe that their audience is more likely to believe that they have been persuaded by the information. Alternatively, they may predict that social rewards or sanctions associated with being perceived as intolerant are not linear in the probability that one is intolerant: for example, they may believe that as long as it appears that is some small probability that they are not intolerant (i.e. because they were exposed to the study and may have been persuaded), their audience will refrain from socially sanctioning them — "innocent until proven guilty". However, to preserve analytic tractability and convey our intuition as simply as possible, we do not formally model either of these channels in our theoretical framework.

Taken together, our evidence suggests a quantitatively important role of the excuse effect (third-order beliefs) in shaping publicly observed behavior relative to the effects of persuasion (first-order beliefs) and the effects of expected persuasion of one's audience (second-order beliefs).

Heterogeneity by local vote shares Which subgroups are driving these effects? Our model predicts that the audience's composition — the share of tolerant vs. intolerant agents — should affect donation decisions by changing the perceived judgment associated with donating. Because we informed respondents that we would promote the website (on which their individual donation decision would be posted) within their geographical area, we might expect that (controlling for the respondent's own private beliefs), respondents in areas with a greater fraction of Republicans should be less sensitive to the availability of a justification than respondents in areas with a lower fraction of Republicans, since Republicans are unlikely to disapprove of the decision to donate to Fund the Wall even in the absence of a justification. We thus pre-registered investigating heterogeneity by the 2016 Republican vote share of respondents' county, which we do by estimating the following specification:

$$\begin{aligned} y_i = \beta_0 &+ \beta_1 \text{Excuse}_i + \beta_2 \text{Control}_i + \beta_3 \text{Excuse}_i \times \text{Rep share}_i + \beta_4 \text{Control}_i \times \text{Rep share}_i \\ &+ \beta_5 \text{Rep share}_i + \varepsilon_i \end{aligned} \tag{10}$$

Table 4 displays the results, revealing striking heterogeneity by the Republican vote share of respondents' counties. In particular, effects are significantly larger for people from counties with a lower Republican vote share, consistent with our model's prediction that the excuse effect should be larger when the share of agents who privately approve of the action is smaller.

5 Robustness

5.1 Demand effects

One concern with our estimated treatment effects could be that respondents across the different treatment conditions hold different beliefs about the experimenter's expectations, and that those beliefs drive the estimated treatment effects. This concern is particularly salient in Experiment 2, which required more explicit instructions in order to fix first- and second-order beliefs. Despite recent research highlighting that respondents are not very elastic to explicit signals of the experimenter's expectations in online surveys (de Quidt et al., 2018), suggesting a limited quantitative importance of demand effects in the context of our experiment, we conduct a number of additional exercises to address the potential for demand effects to bias our findings.

Perceived purpose: machine learning We measured respondent's beliefs about the purpose of the experiment at the end of both Experiment 1 and Experiment 2 using an open-ended question: "If you had to guess, what was the purpose of this study?". To examine whether people in the different treatment conditions hold different beliefs about the purpose of the study, we employ a new approach applying machine learning techniques to these text responses. In particular, we train support vector machine classifiers to predict treatment status given the participant's response. Employing 75 percent of our sample as a training set and the remaining 25 percent as a test set, we show that we cannot predict treatment status in Experiment 1 better than chance (Table A5). Similarly, we cannot predict treatment status better than chance when distinguishing between the Excuse and No Excuse conditions in Experiment 2 (Table A11). However, we show that we can predict assignment to the Control condition better than chance (Table A11) using the text data, which highlights that respondents in the Control condition hold different beliefs from respondents in the Excuse and No Excuse condition. We view this result as validation for our method, as it shows that we would in principle detect differences in perceived purpose between Excuse and No Excuse if such differences were present.

Perceived purpose: hand-coding We also hired two independent research assistants to hand-code the responses to the open-ended purpose question in Experiment 2. Table A9 in the Appendix shows that the majority of our respondents believed that we wanted to study the effects of information on anti-immigrant sentiment or participant's willingness to have their decisions posted on the website. Fewer than 1 percent of our sample correctly guessed the true purpose of our experiment (Column 1). Table A9 also shows that on almost all of the dimensions we code, beliefs about the purpose of the study do not significantly differ between Excuse and No Excuse. The exception is Public Image (Column 3): respondents in the Excuse condition are 2 percentage points more likely than respondents in the No Excuse condition to believe that the study was about whether people were willing to publicly express political views ($p = 0.038$). Although

statistically significant, this different is small in magnitude and cannot explain our effect sizes. Reassuringly, respondents were no more likely to believe that the experimenters were biased in the Excuse condition relative to the No Excuse condition (Column 6, $p = 0.994$).

As suggested by the results of the machine learning exercise described previously, we do find significant differences in perceived purpose between the Control condition and the No Excuse condition, and between the Control condition and the Excuse condition. This is likely due to the fact that we provided respondents in the No Excuse and Excuse conditions information suggesting that undocumented immigrants commit more crimes than US citizens (i.e. the Lott study), while we did not provide any such information to respondents in the Control condition. However, the differences are relatively small, and they do not affect our main comparison of interest (No Excuse vs. Excuse).

Experiment 2 heterogeneity Finally, heterogeneous treatment effects by the county-level Republican vote share are inconsistent with experimenter demand effects driving our findings. In particular, for demand effects alone to explain our findings, we would require that (1) Republicans in counties with a lower Republican vote share are substantially more affected by experimenter demand effects than Republicans in counties with a higher Republican vote share, and (2) that respondents fail to report differences in perceived purpose at the end of the study, a contingency we view as implausible.

5.2 Differential attrition

Could patterns of differential attrition explain the estimated treatment effects in our data? The attrition rate in Experiment 1 is relatively low (8.6 percent). Moreover, attrition rates are virtually identical among respondents in the Excuse and No Excuse conditions ($p = 0.23$) and neither political affiliation nor any other demographic variable systematically predicts differential attrition across treatment arms (Table A4 in the Appendix).

Similarly, in Experiment 2, we find no differential attrition among respondents in the Excuse versus No Excuse condition ($p = 0.47$), and once again, there is no evidence of differential attrition between subgroups (Table A10 in the Appendix). Our base rate of attrition is higher than in Experiment 1 (14.4 percent), which we attribute to the greater length of the survey, but is still relatively low. We do find a small and precisely estimated difference in attrition among respondents in the Control condition (11.6 percent) compared to respondents in the Excuse condition (15.8 percent, $p < 0.001$) and No Excuse condition (15.1 percent, $p < 0.001$), which we again attribute to the greater length of the Excuse and No Excuse versions of the survey.

6 Discussion

In this section, we apply our theoretical framework to shed light on the mechanisms underlying populist rhetoric and incentives in education, and we discuss implications for preventing the spread of misinformation on social media.

6.1 Populist Rhetoric and Dog-Whistling

Müller (2016) argues that populist rhetoric is often characterized by appeals to the beliefs or desires of the "people" or a "silent majority" — a group which often has little to no basis in fact. As Norris and Inglehart (2019) argue:

> "Populist rhetoric tells a simple story about the silent majority of ordinary, hard-working people...Populism rejects the legitimacy of authority derived from scientific evidence, book learning, and reasoned deliberation. Instead, the discourse celebrates the authenticity of direct experience ('Believe me'), mass opinions ('Many people say...'), and quick applause lines ('Build the Wall')."

Johnson (2016) highlights Donald Trump's tendency to use phrases such as "People say..." when discussing politically sensitive issues.[16] (As Rosenblum and Muirhead (2019) argue, this practice is common to a many prominent populist politicians around the world spanning the ideological spectrum.) Through our framework, we interpret this practice two ways. First, populists generates an excuse for his own rhetoric: if their comments backfires, they can distance themselves by claiming they only meant to report on others' beliefs, not express their private beliefs.[17] For example, upon facing criticism for citing a conspiracy theory concerning the suicide of a Vince Foster, a friend of Hillary and Bill Clinton, Trump stated:

> "Somebody asked me the question the other day, and I said that a lot of people are very skeptical as to what happened and how he died. I know nothing about it...I don't think that it's something that should be part of the campaign."

Second, such rhetoric also generates the perception of *common knowledge* of the excuse: by implying that fringe conspiracy theories are known to a large group of people (and appearing to

[16] As of December 1, 2019, Donald Trump had used the phrases *"people say"* or *"people are saying"* a combined 591 times in speeches, interviews, and Tweets since launching his presidential campaign in June 2015 (Factbase, 2019). Johnson (2016) highlights a number of examples, including comments about Barack Obama's religion, Ted Cruz' birthplace, the Iran nuclear deal, extremist Islamic terrorism, and the existence of a "deep state" dedicated to undermining the Trump administration.

[17] This is an example of a *motte-and-bailey* argument (Shackel, 2005), through which a speaker can make a controversial, difficult-to-defend statement (the "bailey") and then retreat to a less controversial and easier-to-defend position when challenged (the "motte"), claiming that the audience had misinterpreted his or her original statement. Such rhetorical devices can thus be used to generate excuses for making false, extreme, or misleading remarks.

endorse the theory themselves), populists' comments positively update their audience's perceptions as to the likelihood that others will believe that they themselves were convinced by the story, and through this channel, the perceived effectiveness of the excuse.

A closely related phenomenon is dog-whistling: "sending a message to certain potential supporters in such a way as to make it inaudible to others whom it might alienate or deniable for still others who would find any explicit appeal along those lines offensive" (Goodin and Saward, 2005). The term was first introduced in its modern form to describe the rhetoric of Australian Prime Minister John Howard over the course of his tenure (1996-2007) and re-election campaigns, during which opponents accused him of employing terms such as "un-Australian" and "mainstream" to signal racist positions while maintaining plausible deniability (Soutphommasane, 2009); the term has also been used retrospectively to describe the Republican Party's "Southern strategy" to win white support in the South by appealing to racist tensions during the 1970s and 80s (Haney-López, 2014). In a 1981 interview, Republican strategist and Republican National Committee chairman Lee Atwater described the strategy as follows:

> "Y'all don't quote me on this. You start out in 1954 by saying, 'N—, n—, n—.' By 1968 you can't say 'n—': that hurts you. Backfires. So you say stuff like forced busing, states' rights and all that stuff. You're getting so abstract now [that] you're talking about cutting taxes, and all these things you're talking about are totally economic things and a byproduct of them is [that] blacks get hurt worse than whites. And subconsciously maybe that is part of it. I'm not saying that." (Lamis, ed, 1999)

As with "people say" and related language, "dog-whistles" generate two types of excuses: one for the politician vis-a-vis the public, and one for the politician's supporters vis-a-vis others who disapprove of the statement, allowing them to publicly support the politician and his or her policies without incurring social stigma.[18]

6.2 Educational Incentives and "Acting White"

Austen-Smith and Fryer (2005) model negative stigma associated with educational investment – "Acting White". This stigma might arise if choosing to study signals that individual has limited social opportunities (i.e. the opportunity cost of studying is lower), or if choosing to study signals that the individual is weakly attached to his social group (e.g. because he or she hopes to do well enough academically to leave the community). Bursztyn et al. (2019a) show that students are willing to forego a valuable SAT preparatory package when their decision to take it up is public to their peers, and they demonstrate that in a lower-achieving school with a greater fraction of minority students, this effect is due to students' desire to avoid signaling effort to their peers. As

[18]Indeed, a third type of excuse may be a "self-excuse" for politician's supporters who do not want to admit to themselves that they endorse racist positions, as in Bénabou and Tirole (2011).

the authors point out, in such settings, providing monetary incentives for exerting educational effort (as discussed in Levitt et al. 2016) might provide students with an excuse, allowing them to claim that they are studying not because they care about their grades but rather because they desire the monetary reward. For similar reasons, in these settings, cold-calling after asking a question might be preferable to allowing students to volunteer answers.

6.3 Debunking Fake News

Our findings also have implications for debunking fake news spread online and offline. In particular, our model of excuses suggests that in order to prevent a given fake news story from spreading, it is insufficient to debunk it *privately*; instead, it is crucial to generate common knowledge that the excuse is invalid. This insight has valuable implications for institutional policy. Among other platforms, Facebook has been experimenting with warning users before they post an article flagged as fake news. Facebook also currently flags fake or misleading news when it appears on users' timelines (e.g. because a friend shared it). The former initiative maps closely onto a "first-order" debunking in our model (private persuasion), while the second initiative maps onto a "second-order" debunking (debunking one's audience). Yet to the extent that Facebook does not yet debunk all users, it generates a ready-made excuse for sharing fake news: posters can credibly claim that they were not warned the news was fake. Our results suggest it is important not only to debunk both the poster and the audience, but also to make it clear to the poster that the audience will *know that he or she was debunked before posting*. This could be done by including a screenshot in the pre-post warning shown to the poster of what his or her post will look like to others, in which the sentence "The poster was warned that this link has been flagged as fake or misleading before posting" is clearly visible. An alternative and simpler path would be to simply roll out the feature to the entire user-base, generating common knowledge that all users are warned before posting fake news. Because the general equilibrium results of such a change differ significantly from the partial equilibrium results by altering the structure of third-order beliefs, current estimates of the effects of debunking on users' propensity to share fake news may substantially understate the true effects that would be realized if platforms were to scale up the feature to all users.

7 Conclusion

Motivated by the recent wave of anti-immigrant sentiment in the United States and Western Europe in general, and by the campaign rhetoric of Donald Trump in particular, we study the effect of a common narrative — that immigrants commit crimes at vastly higher rates than citizens — on respondent's willingness to donate to an anti-immigrant organization. We show that the availability of this excuse reduces the extent to which those who disapprove of the organization judge the agent for donating, and that excuses can thus increase the prevalence of stigmatized behavior in equilibrium.

Our results suggest several directions for further research. First, to what extent can dramatic changes in public behavior — for example, waves of violence against minority groups or revolutions against authoritarian regimes — be attributed to the availability of excuses? Moreover, what implications does this phenomenon have for the "supply side" of excuses: can "excuse entrepreneurs" who are able to generate common knowledge of plausible justifications to act in a potentially stigmatized manner cause striking reversals of social norms, even if their persuasive impact is limited, and can similar patterns help explain the rising popularity of ideologically extreme media outlets? Finally, to what extent can excuses, by becoming socially widespread and through mere repetition, grow privately persuasive?

References

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya, "Radio and the Rise of The Nazis in Prewar Germany," *The Quarterly Journal of Economics*, November 2015, *130* (4), 1885–1939.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva, "Immigration and redistribution," Technical Report, National Bureau of Economic Research 2018.

Andreoni, James and B. Douglas Bernheim, "Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica*, September 2009, *77* (5), 1607–1636.

Ariely, Dan, Anat Bracha, and Stephan Meier, "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, March 2009, *99* (1), 544–555.

Austen-Smith, David and Roland G. Fryer, "An Economic Analysis of "Acting White"," *The Quarterly Journal of Economics*, May 2005, *120* (2), 551–583.

Banerjee, Abhijit, Eliana La Ferrara, and Victor Orozco, "The entertaining way to behavioral change," *Unpublished manuscript*, 2019.

Bénabou, Roland and Jean Tirole, "Incentives and Prosocial Behavior," *American Economic Review*, December 2006, *96* (5), 1652–1678.

_ and _, "Identity, Morals, and Taboos: Beliefs as Assets," *The Quarterly Journal of Economics*, May 2011, *126* (2), 805–855.

_ , Armin Falk, and Jean Tirole, "Narratives, imperatives, and moral reasoning," Technical Report, National Bureau of Economic Research 2018.

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott, "Misperceived Social Norms: Female Labor Force Participation in Saudi Arabia," Working Paper 24736, National Bureau of Economic Research June 2018.

_ , Georgy Egorov, and Robert Jensen, "Cool to Be Smart or Smart to Be Cool? Understanding Peer Pressure in Education," *The Review of Economic Studies*, July 2019, *86* (4), 1487–1526.

_ , _ , and Stefano Fiorin, "From Extreme to Mainstream: How Social Norms Unravel," *NBER Working Paper Series*, May 2019.

_ , _ , Ruben Enikolopov, and Maria Petrova, "Social media and xenophobia: evidence from Russia," Technical Report, National Bureau of Economic Research 2020.

_ , Thomas Fujiwara, and Amanda Pallais, "'Acting Wife': Marriage Market Incentives and Labor Market Investments," *American Economic Review*, November 2017, *107* (11), 3288–3319.

Cantoni, Davide, David Y. Yang, Noam Yuchtman, and Y. Jane Zhang, "Protests as Strategic Games: Experimental Evidence from Hong Kong's Antiauthoritarian Movement," *The Quarterly Journal of Economics*, May 2019, *134* (2), 1021–1077.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang, "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory*, October 2007, *33* (1), 67–80.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth, "Measuring and Bounding Experimenter Demand," *American Economic Review*, November 2018, *108* (11), 3266–3302.

DellaVigna, Stefano and Eliana La Ferrara, "Economic and social impacts of the media," in "Handbook of media economics," Vol. 1, Elsevier, 2015, pp. 723–768.

_ , John A. List, Ulrike Malmendier, and Gautam Rao, "Voting to Tell Others," *The Review of Economic Studies*, January 2017, *84* (1), 143–181.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova, "Social Media and Protest Participation: Evidence from Russia," SSRN Scholarly Paper ID 2696236, Social Science Research Network, Rochester, NY November 2019.

_ and Maria Petrova, "Media capture: empirical evidence," in "Handbook of media Economics," Vol. 1, Elsevier, 2015, pp. 687–700.

Exley, Christine L., "Excusing Selfishness in Charitable Giving: The Role of Risk," *The Review of Economic Studies*, April 2016, *83* (2), 587–628.

Factbase, "Donald Trump Unabridged - Speeches, Tweets, Interviews," https://factba.se 2019.

Ferrara, Eliana La, "Mass media and social change: Can we use television to fight poverty?," *Journal of the European Economic Association*, 2016, *14* (4), 791–827.

_ , Alberto Chong, and Suzanne Duryea, "Soap operas and fertility: Evidence from Brazil," *American Economic Journal: Applied Economics*, 2012, *4* (4), 1–31.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as Data," Technical Report w23276, National Bureau of Economic Research, Cambridge, MA March 2017.

Goodin, Robert E. and Michael Saward, "Dog Whistles and Democratic Mandates," *The Political Quarterly*, 2005, *76* (4), 471–476.

Grigorieff, Alexis, Christopher Roth, and Diego Ubfal, "Does Information Change Attitudes Towards Immigrants?," *Demography*, 2020.

Hager, Anselm, Lukas Hensel, Johannes Hermle, and Christopher Roth, "Political Activists as Free-Riders: Evidence from a Natural Field Experiment," *IZA discussion paper*, 2019.

Hamman, John R., George Loewenstein, and Roberto A. Weber, "Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship," *American Economic Review*, September 2010, *100* (4), 1826–1846.

Haney-López, Ian, *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*, Oxford ; New York: Oxford University Press, 2014.

Hopkins, Daniel J, John Sides, and Jack Citrin, "The muted consequences of correct information about immigration," *The Journal of Politics*, 2019, *81* (1), 315–320.

Johnson, Jenna, "'A Lot of People Are Saying . . . ': How Trump Spreads Conspiracies and Innuendoes," *Washington Post*, June 2016.

Karing, Anne, "Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone," *University of California, Berkeley*, 2018.

Kuran, Timur, *Private Truths, Public Lies: The Social Consequences of Preference Falsification*, 1st harvard university press pbk. ed ed., Cambridge, Mass: Harvard University Press, 1997. OCLC: ocm40097352.

Kuziemko, Ilyana, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva, "How elastic are preferences for redistribution? Evidence from randomized survey experiments," *American Economic Review*, 2015, *105* (4), 1478–1508.

Lamis, Alexander P., ed., *Southern Politics in the 1990s*, Baton Rouge: Louisiana State University Press, 1999.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber, "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics*, January 2012, *4* (1), 136–163.

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff, "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance," *American Economic Journal: Economic Policy*, November 2016, *8* (4), 183–219.

Lott, John R., "Undocumented Immigrants, U.S. Citizens, and Convicted Criminals in Arizona," SSRN Scholarly Paper ID 3099992, Social Science Research Network, Rochester, NY February 2018.

Müller, Jan-Werner, *What Is Populism?*, Philadelphia: University of Pennsylvania Press, 2016. OCLC: ocn946693516.

Müller, Karsten and Carlo Schwarz, "Making America Hate Again? Twitter and Hate Crime Under Trump," SSRN Scholarly Paper ID 3149103, Social Science Research Network, Rochester, NY March 2018.

Norris, Pippa and Ronald Inglehart, *Cultural Backlash: Trump, Brexit, and Authoritarian Populism*, New York, NY: Cambridge University Press, February 2019.

Perez-Truglia, Ricardo and Guillermo Cruces, "Partisan Interactions: Evidence from a Field Experiment in the United States," *Journal of Political Economy*, August 2017, *125* (4), 1208–1243.

Pew, "Most Border Wall Opponents, Supporters Say Shutdown Concessions Are Unacceptable," January 2019.

Rosenblum, Nancy L. and Russell Muirhead, *A Lot of People Are Saying: The New Conspiracism and the Assault on Democracy*, Princeton, New Jersey: Princeton University Press, April 2019.

Sessions, Jeff, "Attorney General Sessions Delivers Remarks on National Security and Immigration Priorities of the Administration," https://www.justice.gov/opa/speech/attorney-general-sessions-delivers-remarks-national-security-and-immigration-priorities January 2018.

Shackel, Nicholas, "The Vacuity of Postmodernist Methodology," *Metaphilosophy*, 2005, *36* (3), 295–320.

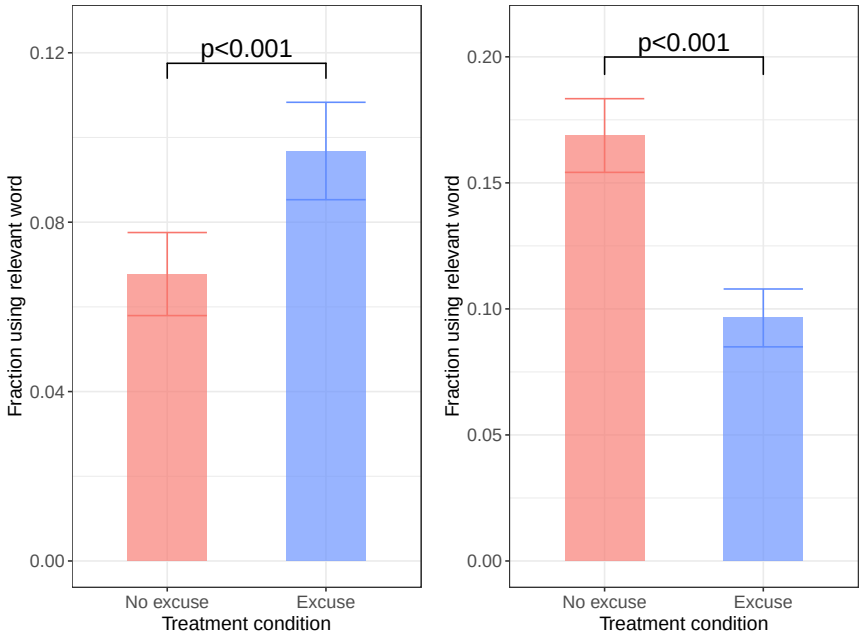
Soutphommasane, Tim, *Reclaiming Patriotism: Nation-Building for Australian Progressives* Australian Encounters, Cambridge ; New York : [Clayton, Vic.]: Cambridge University Press ; Monash University, National Centre for Australian Studies, 2009. OCLC: ocn320818181.

Wood, Thomas and Ethan Porter, "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence," *Political Behavior*, 2019, *41* (1), 135–163.

Yanagizawa-Drott, David, "Propaganda and Conflict: Evidence from the Rwandan Genocide," *The Quarterly Journal of Economics*, November 2014, *129* (4), 1947–1994.

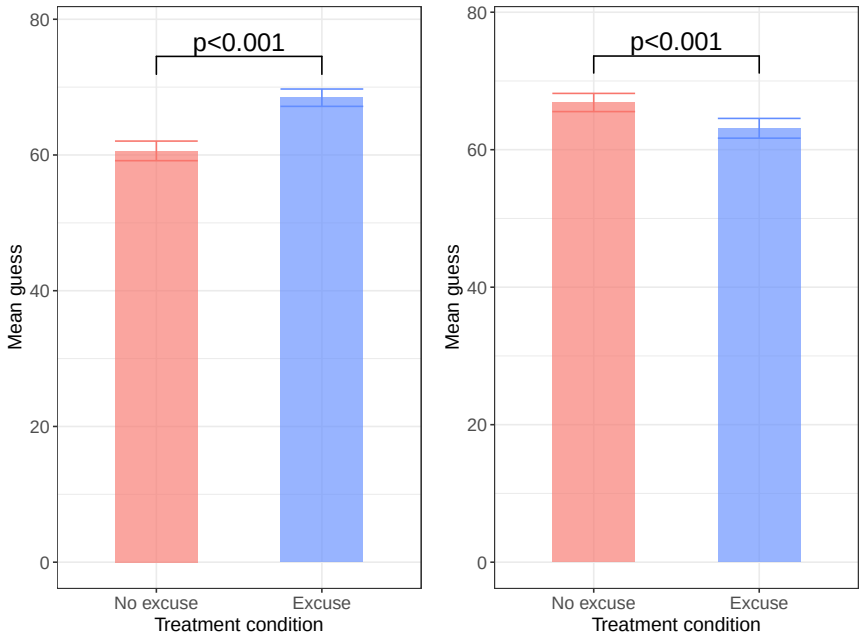
Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov, "Political Effects of the Internet and Social Media," *Forthcoming, Annual Review of Economics. DOI/10.1146/annurev-economics-081919-050239*, 2019.

Figures



(a) Gullibility: word count

(b) Bias: word count



(c) Gullibility: Scale

(d) Bias: Scale

Figure 1: Experiment 1: Type inference based on text analysis and scales

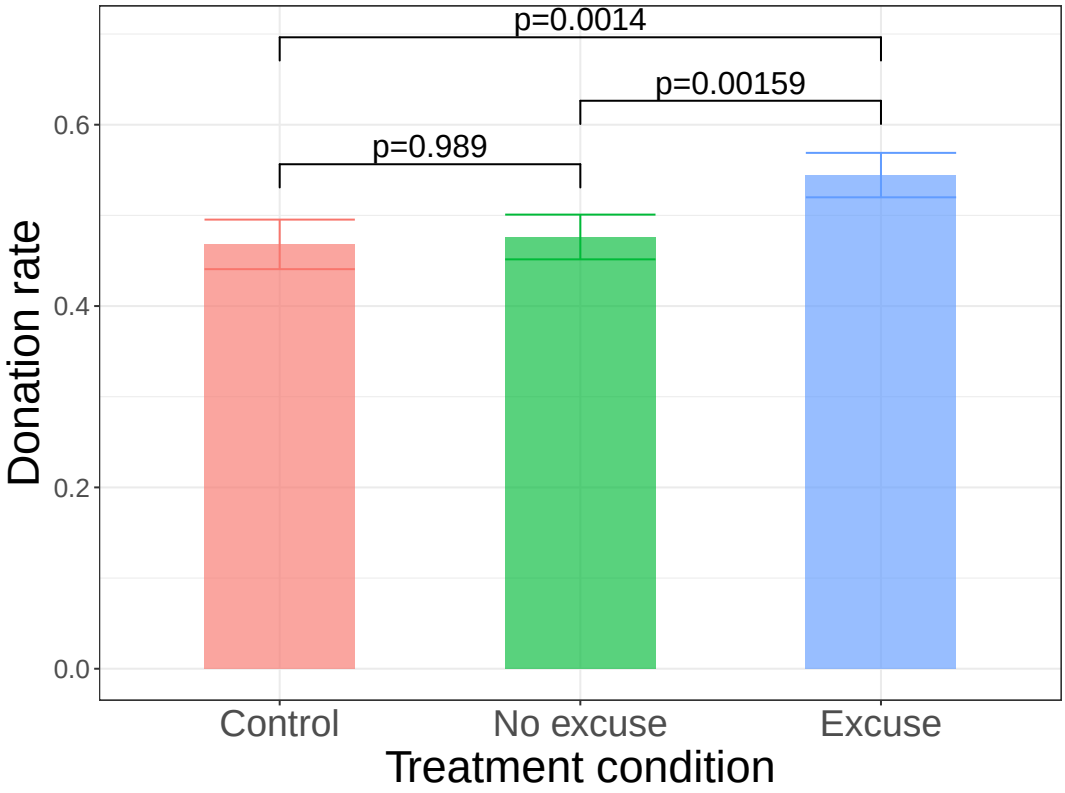
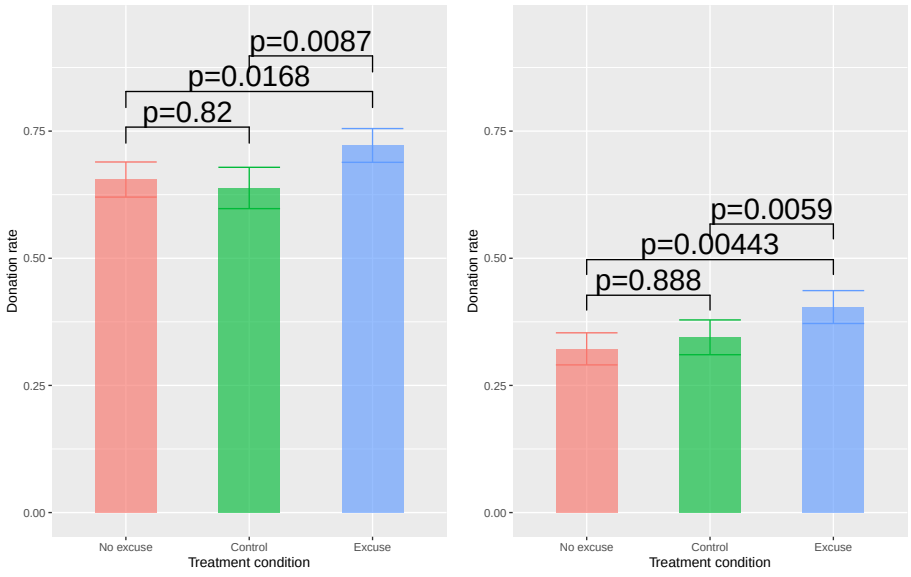


Figure 2: Experiment 2: Donation rates by group



(a) Republicans

(b) Independents

Figure 3: Experiment 2: Donation rates by group and partisan affiliation

Tables

Table 1: Experiment 2: Inferred donation motives

	Dependent variable:					
	Inference about partner's donation motive					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: *Probability of using word relating to bias*						
Excuse	-0.070^{***}	-0.068^{***}	-0.068^{***}	-0.072^{***}	-0.072^{***}	-0.072^{***}
	(0.012)	(0.012)	(0.012)	(0.009)	(0.009)	(0.009)
Constant	0.172^{***}	0.310^{***}	0.291^{***}	0.169^{***}	0.288^{***}	0.268^{***}
	(0.009)	(0.071)	(0.072)	(0.007)	(0.055)	(0.055)
Observations	3,047	3,047	3,047	5,065	5,065	5,065
R^2	0.010	0.028	0.029	0.011	0.024	0.025
Adjusted R^2	0.010	0.024	0.024	0.011	0.021	0.022
Panel B: *Probability of using word relating to gullibility*						
Excuse	0.031^{***}	0.032^{***}	0.032^{***}	0.029^{***}	0.029^{***}	0.029^{***}
	(0.010)	(0.010)	(0.010)	(0.008)	(0.008)	(0.008)
Constant	0.069^{***}	0.112^{*}	0.118^{**}	0.068^{***}	0.111^{**}	0.113^{**}
	(0.007)	(0.058)	(0.058)	(0.005)	(0.045)	(0.045)
Observations	3,047	3,047	3,047	5,065	5,065	5,065
R^2	0.003	0.010	0.011	0.003	0.009	0.009
Adjusted R^2	0.003	0.006	0.006	0.003	0.006	0.006
Demographic controls	No	Yes	Yes	No	Yes	Yes
Partisan affiliation controls	No	No	Yes	No	No	Yes
Include pilot data	No	No	No	Yes	Yes	Yes

Notes: The dependent variable in Panel A is the probability that the respondent uses a word relating to bias when describing why he or she thinks the matched respondent donated to Fund the Wall, while the dependent variable in Panel B is the probability that the respondent uses a word relating to gullibility in response to the same question. Columns 1-3 report results estimated on the sample from the main experiment, while Columns 4-6 pool the sample from the main experiment with the sample from the pilot. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Partisan affiliation controls include a dummy for strong Democrats, with weak Democrats as the reference category. Robust standard errors are reported.

Table 2: Experiment 2: Inferred bias and gullibility scores

	Dependent variable:					
	Inference about partner's score					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: *Bias (z-score)*						
Excuse	-0.134^{***}	-0.133^{***}	-0.133^{***}	-0.149^{***}	-0.152^{***}	-0.152^{***}
	(0.051)	(0.051)	(0.051)	(0.040)	(0.039)	(0.039)
Constant	0.058	-0.159	-0.174	0.074^{***}	0.047	0.006
	(0.036)	(0.282)	(0.285)	(0.028)	(0.224)	(0.226)
Observations	1,524	1,524	1,524	2,532	2,532	2,532
R^2	0.004	0.038	0.038	0.006	0.037	0.037
Adjusted R^2	0.004	0.030	0.029	0.005	0.032	0.032
Panel B: *Gullibility (z-score)*						
Excuse	0.321^{***}	0.312^{***}	0.310^{***}	0.313^{***}	0.316^{***}	0.317^{***}
	(0.050)	(0.050)	(0.050)	(0.039)	(0.039)	(0.039)
Constant	-0.155^{***}	0.016	-0.100	-0.159^{***}	-0.116	-0.224
	(0.036)	(0.299)	(0.301)	(0.028)	(0.231)	(0.233)
Observations	1,523	1,523	1,523	2,533	2,533	2,533
R^2	0.026	0.060	0.065	0.025	0.055	0.059
Adjusted R^2	0.025	0.052	0.056	0.024	0.050	0.053
Demographic controls	No	Yes	Yes	No	Yes	Yes
Partisan affiliation controls	No	No	Yes	No	No	Yes
Include pilot data	No	No	No	Yes	Yes	Yes

Notes: The dependent variable in Panel A is the negative of the z-score of the respondent's guess as to his or her matched respondent's score on the Foreign Culture Tolerance Scale, where we take the negative to interpret higher values as greater bias. The dependent variable in Panel B is the z-score of the respondent's guess as to his or her matched respondent's score on the Gullibility Scale. Both scales were originally scored between 0 and 100. Columns 1-3 report results estimated on the sample from the main experiment, while Columns 4-6 pool the sample from the main experiment with the sample from the pilot. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. The partisan affiliation control is an indicator that takes value 1 if the respondent self-reports being a strong Democrat, with weak Democrat as the reference category. Robust standard errors are reported.

Table 3: Experiment 2: Main results

	Dependent variable:					
	Donated to Fund the Wall					
	(1)	(2)	(3)	(4)	(5)	(6)
Excuse	0.064^{***}	0.060^{***}	0.065^{***}	0.068^{***}	0.067^{***}	0.074^{***}
	(0.020)	(0.020)	(0.018)	(0.018)	(0.017)	(0.016)
Control	-0.0003	-0.007	-0.004	-0.008	-0.011	-0.00004
	(0.020)	(0.019)	(0.018)	(0.019)	(0.018)	(0.017)
p-value (Excuse = Control)	0.0014	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Demographic controls	No	Yes	Yes	No	Yes	Yes
Partisan affiliation controls	No	No	Yes	No	No	Yes
Include pilot data	No	No	No	Yes	Yes	Yes
DV mean	0.489	0.489	0.489	0.498	0.498	0.498
DV std. dev.	0.500	0.500	0.500	0.500	0.500	0.500
Observations	3,728	3,728	3,728	4,444	4,432	4,432
R^2	0.004	0.060	0.188	0.005	0.061	0.198
Adjusted R^2	0.003	0.056	0.184	0.004	0.058	0.195

Notes: Columns 1-3 report results estimated on the sample from the main experiment, while Columns 4-6 pool the sample from the main experiment with the sample from the pilot. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Partisan affiliation controls include dummies for strong Republican, weak Republican, Republican-leaning Independent, and Democrat-leaning Independent. Robust standard errors are reported.

Table 4: Experiment 2: County heterogeneity

	Dependent variable:					
	Donated to Fund the Wall					
	All		Republicans		Independents	
	(1)	(2)	(3)	(4)	(5)	(6)
Excuse	0.062^{***}	0.070^{***}	0.061^{**}	0.069^{***}	0.061^{**}	0.071^{***}
	(0.019)	(0.016)	(0.029)	(0.025)	(0.024)	(0.022)
Excuse × County Republican vote share	-0.037^{**}	-0.037^{**}	-0.011	-0.013	-0.059^{**}	-0.059^{***}
	(0.018)	(0.016)	(0.029)	(0.025)	(0.024)	(0.022)
Control	-0.005	-0.002	-0.006	-0.004	0.005	0.008
	(0.018)	(0.017)	(0.028)	(0.026)	(0.024)	(0.023)
Control × County Republican vote share	0.019	0.017	0.018	0.012	0.021	0.022
	(0.013)	(0.012)	(0.020)	(0.017)	(0.018)	(0.016)
County Republican vote share	0.010	0.011	-0.008	-0.006	0.026	0.023
	(0.018)	(0.017)	(0.029)	(0.027)	(0.024)	(0.023)
Demographic controls	Yes	Yes	Yes	Yes	Yes	Yes
Partisan affiliation controls	Yes	Yes	Yes	Yes	Yes	Yes
Include pilot data	No	Yes	No	Yes	No	Yes
DV mean	0.489	0.498	0.489	0.498	0.489	0.498
DV std. dev.	0.500	0.500	0.500	0.500	0.500	0.500
Observations	3,631	4,315	1,551	1,920	2,080	2,395
R^2	0.192	0.203	0.071	0.073	0.142	0.156
Adjusted R^2	0.188	0.199	0.060	0.064	0.134	0.150

Notes: Republican vote share is scaled to a standard normal distribution. Columns 1-2 include both Independents and Republicans, Columns 3-4 limit the sample to Republicans, and Columns 5-6 limit the sample to Independents. Columns 1, 3, and 5 report results estimated on the sample from the main experiment, while Columns 2, 4, and 6 pool the sample from the main experiment with the sample from the pilot. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Partisan affiliation controls include dummies for strong Republican, weak Republican, Republican-leaning Independent, and Democrat-leaning Independent. Robust standard errors are reported.

Appendices

Supplementary Appendix: not for publication

A Appendix Figures and Tables

Table A4: Experiment 1: Attrition

	Dependent variable:
	Attrited
	(1)
Age	0.001
	(0.002)
Age squared	0.00001
	(0.00002)
Black	−0.017
	(0.025)
Asian	−0.010
	(0.032)
White	−0.038*
	(0.022)
Hispanic	0.006
	(0.016)
Male	−0.029***
	(0.010)
High school	0.0004
	(0.039)
Some college, no degree	−0.023
	(0.038)
Associate degree	−0.030

Continued on next page

Table A4 – *Continued from previous page*

	Attrited (1)
	(0.040)
Bachelor degree	−0.034
	(0.038)
Post-bachelor degree	−0.053
	(0.040)
Strong Democrat	−0.013
	(0.011)
Excuse × Age	−0.001
	(0.002)
Excuse × Age squared	0.00001
	(0.00002)
Excuse × Black	0.046
	(0.034)
Excuse × Asian	−0.013
	(0.044)
Excuse × White	0.045
	(0.030)
Excuse × Hispanic	0.030
	(0.023)
Excuse × Male	−0.016
	(0.015)
Excuse × High school	−0.008
	(0.048)
Excuse × Some college, no degree	−0.008
	(0.047)
Excuse × Associate degree	−0.023
	(0.050)
Excuse × Bachelor degree	−0.010

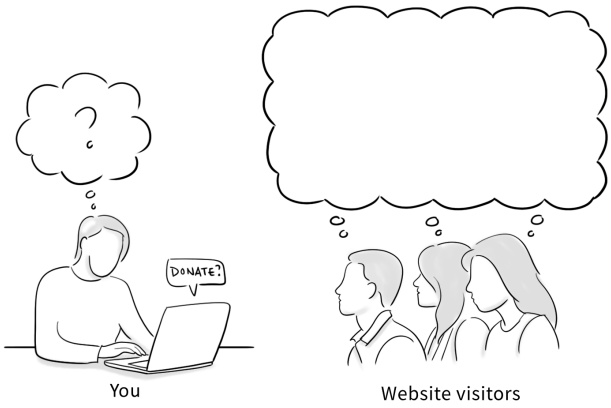
Continued on next page

Table A4 – *Continued from previous page*

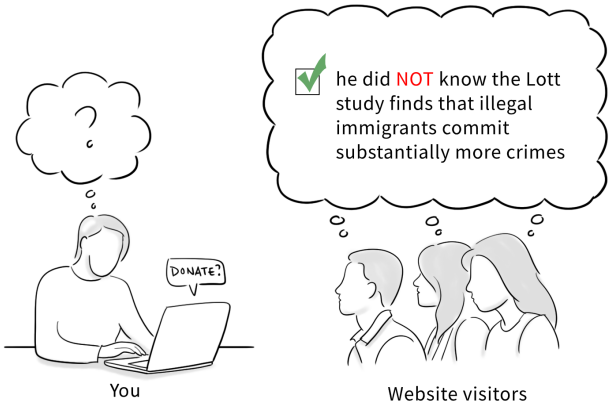
	Attrited (1)
	(0.048)
Excuse × Post-bachelor degree	−0.010
	(0.051)
Excuse × Strong Democrat	−0.001
	(0.016)
DV mean (no excuse)	0.086
DV mean (excuse)	0.077
Observations	5,515
R^2	0.015
Adjusted R^2	0.010

Notes: The dependent variable is an indicator that takes value 1 if the respondent attrited post-randomization. Robust standard errors are reported.

Panel A: Control condition



Panel B: No Excuse condition



Panel C: Excuse condition



Figure A1: Illustrations of audience's information set

Table A1: Experiment 1: Sample representativeness

	Experiment 1	Pew
Variables:	(1)	(2)
Age	41.58	46.67
Black	0.18	0.26
Asian	0.05	0.05
White	0.62	0.49
Hispanic	0.14	0.17
Male	0.47	0.39
Bachelors degree or higher	0.46	0.36
Observations	5151	4005

Notes: Mean of respondent characteristics in experiment 1 and the 2018 Pew Research Center's American Trends Panel Wave 39. Attriters dropped from sample.

Table A2: Experiment 1: Balance of covariates

	Overall		Excuse	No Excuse	p-value
	mean	std.dev.	mean	mean	(E=NE)
Variables:	(1)	(2)	(3)	(4)	(5)
Age	41.376	15.639	41.703	41.048	0.247
Black	0.182	0.386	0.186	0.179	0.612
Asian	0.045	0.208	0.049	0.042	0.386
White	0.710	0.454	0.703	0.716	0.455
Hispanic	0.140	0.347	0.136	0.144	0.561
Male	0.450	0.498	0.451	0.448	0.840
High school diploma	0.983	0.130	0.983	0.983	0.998
Bachelors degree	0.446	0.497	0.454	0.439	0.391

Notes: p-values based on robust standard errors reported. Attriters dropped from sample.

Table A3: Experiment 2: Relationship between perceived motive and scores

	Dependent variable:					
	Inference about partner's score					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: *Bias (z-score)*						
Used bias word	0.477***	0.473***	0.473***	0.502***	0.500***	0.500***
	(0.072)	(0.072)	(0.072)	(0.057)	(0.056)	(0.056)
Constant	−0.078***	−0.283	−0.301	−0.070***	−0.107	−0.149
	(0.027)	(0.278)	(0.281)	(0.021)	(0.220)	(0.222)
Observations	1,524	1,524	1,524	2,532	2,532	2,532
R^2	0.028	0.061	0.061	0.030	0.060	0.061
Adjusted R^2	0.027	0.053	0.052	0.030	0.056	0.056
Panel B: *Gullibility (z-score)*						
Used gullibility word	0.520***	0.486***	0.484***	0.429***	0.398***	0.397***
	(0.089)	(0.088)	(0.088)	(0.071)	(0.070)	(0.070)
Constant	−0.039	0.064	−0.055	−0.037*	−0.003	−0.107
	(0.026)	(0.300)	(0.302)	(0.021)	(0.232)	(0.234)
Observations	1,523	1,523	1,523	2,533	2,533	2,533
R^2	0.022	0.054	0.059	0.014	0.042	0.046
Adjusted R^2	0.021	0.046	0.051	0.014	0.037	0.041
Demographic controls	No	Yes	Yes	No	Yes	Yes
Partisan affiliation controls	No	No	Yes	No	No	Yes
Include pilot data	No	No	No	Yes	Yes	Yes

Notes: The dependent variable in Panel A is the negative of the z-score of the respondent's guess as to his or her matched respondent's score on the Foreign Culture Tolerance Scale, where we take the negative to interpret higher values as greater bias. The dependent variable in Panel B is the z-score of the respondent's guess as to his or her matched respondent's score on the Gullibility Scale. Both scales were originally scored between 0 and 100. Columns 1-3 report results estimated on the sample from the main experiment, while Columns 4-6 pool the sample from the main experiment with the sample from the pilot. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. The partisan affiliation control is an indicator that takes value 1 if the respondent self-reports being a strong Democrat, with weak Democrat as the reference category. Robust standard errors are reported.

Table A5: Experiment 1: Condition prediction confusion matrix

	Predicted Excuse	**Predicted Excuse**
True Excuse	231	166
True No Excuse	206	176
		Overall accuracy: 0.5225

Notes: Each cell reports the number of individuals who were assigned to the condition (Excuse or No Excuse) in the corresponding row and who were classified by the Support Vector Machine as belonging to the condition in the corresponding column. The classifier was trained on a 75% sample of the data; the table reports prediction results on the test set of the remaining 25% of the data. Overall accuracy is calculated as the proportion of correct predictions.

Table A6: Experiment 2: Sample representativeness

	Experiment 2	Pew
	(1)	(2)
Panel A: *Republican*		
Age	47.11	49.50
Black	0.03	0.02
Asian	0.03	0.03
White	0.83	0.84
Hispanic	0.09	0.08
Male	0.49	0.51
Bachelors degree or higher	0.38	0.29
Observations	2022	2879
Panel B: *Independent*		
Age	43.53	44.96
Black	0.11	0.08
Asian	0.05	0.04
White	0.69	0.70
Hispanic	0.12	0.13
Male	0.50	0.53
Bachelors degree or higher	0.37	0.34
Observations	2531	2622

Notes: Mean of respondent characteristics in experiment 2 and the 2018 Pew Research Center's American Trends Panel Wave 39. Attriters dropped from sample.

Table A7: Experiment 2: Balance of covariates

	Overall		Excuse	No Excuse	Control	p-values		
	mean	std.dev.	mean	mean	mean	(E=NE)	(E=C)	(NE=C)
Variables:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Age	44.935	15.704	45.100	44.798	44.909	0.633	0.758	0.857
Black	0.076	0.266	0.069	0.088	0.072	0.085	0.797	0.135
Asian	0.043	0.202	0.041	0.042	0.045	0.877	0.657	0.772
White	0.821	0.383	0.826	0.815	0.823	0.465	0.877	0.556
Hispanic	0.110	0.313	0.113	0.107	0.111	0.644	0.854	0.776
Male	0.500	0.500	0.494	0.507	0.498	0.535	0.835	0.674
High school diploma	0.976	0.153	0.977	0.975	0.977	0.821	0.987	0.831
Bachelors degree	0.379	0.485	0.393	0.369	0.375	0.213	0.356	0.734
Republican	0.425	0.494	0.419	0.436	0.420	0.389	0.955	0.413

Notes: p-values based on robust standard errors reported. Attriters dropped from sample.

Table A8: Experiment 2: Party heterogeneity

	Dependent variable:			
	Donated to Fund the Wall			
	Republicans		Independents	
	(1)	(2)	(3)	(4)
Excuse	0.068**	0.079***	0.068***	0.074***
	(0.028)	(0.024)	(0.024)	(0.023)
Control	−0.006	−0.008	0.003	0.013
	(0.028)	(0.026)	(0.024)	(0.024)
p-value (Excuse = Control)	0.0087	0.0011	0.0059	0.0089
Demographic controls	Yes	Yes	Yes	Yes
Partisan affiliation controls	Yes	Yes	Yes	Yes
Include pilot data	No	Yes	No	Yes
DV mean	0.670	0.674	0.356	0.358
DV std. dev.	0.470	0.469	0.479	0.469
Observations	1,582	1,961	2,146	2,471
R^2	0.071	0.049	0.133	0.049
Adjusted R^2	0.062	0.042	0.127	0.043

Notes: Columns 1-2 limit the sample to Republicans, while Columns 3-4 limit the sample to Independents. Columns 1 and 3 report results estimated on the sample from the main experiment, while Columns 2 and 4 pool the sample from the main experiment with the sample from the pilot. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Partisan affiliation controls include dummies for strong Republican, weak Republican, Republican-leaning Independent, and Democrat-leaning Independent. Robust standard errors are reported.

Table A9: Experiment 2: Perceived purpose of study

	Dependent variable:					
	Excuse	Immigration attitudes	Public image	Information	Persuasion	Biased
	(1)	(2)	(3)	(4)	(5)	(6)
Excuse	−0.005	0.009	0.020**	0.012	−0.013	−0.00003
	(0.003)	(0.015)	(0.010)	(0.015)	(0.011)	(0.013)
Control	−0.003	0.133***	0.037***	−0.012	−0.081***	−0.036**
	(0.003)	(0.015)	(0.010)	(0.016)	(0.012)	(0.014)
p-value (Excuse = Control)	0.62	< 0.001	0.098	0.13	< 0.001	0.012
Demographic controls	Yes	Yes	Yes	Yes	Yes	Yes
Partisan affiliation controls	Yes	Yes	Yes	Yes	Yes	Yes
Include pilot data	Yes	Yes	Yes	Yes	Yes	Yes
DV mean	0.007	0.228	0.083	0.240	0.122	0.176
DV std. dev.	0.084	0.419	0.275	0.427	0.327	0.381
Observations	4,514	4,514	4,514	4,513	4,514	4,512
R^2	0.004	0.028	0.018	0.011	0.022	0.009
Adjusted R^2	0.001	0.025	0.015	0.007	0.019	0.006

Notes: The dependent variable in each column is an indicator for whether the respondent's perceived purpose of the study was coded to fall into the corresponding category. "Excuse" takes value 1 if the respondent correctly inferred the study was about whether knowing that others will know one had an "excuse" for donating would affect the donation decision. "Immigration attitudes" takes value 1 if the respondent stated the study was about attitudes toward immigration. "Public image" takes value 1 if the respondent stated the study was about whether knowing one's decision will be observable to others would affect the donation decision. "Information" takes value 1 if the respondent stated the study was about disseminating information about immigration. "Persuasion" takes value 1 if the respondent stated the researchers were attempting to persuade them either to donate or not to donate. "Bias" takes value 1 if the respondent stated the researchers were biased. "Other" takes value 1 if the respondent stated a purpose that did not fall into any of the above categories. Categories other than "Other" are not mutually exclusive. All specifications pool the main experiment and the pilot and contol for demographics and partisan affiliation. Demographic controls include age, age squared, a set of race indicators, a Hispanic indicator, a male indicator, and a set of education indicators. Partisan affiliation controls include dummies for strong Republican, weak Republican, Republican-leaning Independent, and Democrat-leaning Independent. Robust standard errors are reported.

Table A10: Experiment 2: Attrition

	Dependent variable:
	Attrited
	(1)
Age	−0.001
	(0.003)
Age squared	0.00004
	(0.00003)
Black	−0.008
	(0.048)
Asian	0.049
	(0.055)
White	0.036
	(0.040)
Hispanic	−0.027
	(0.030)
Male	−0.056***
	(0.017)
High school	−0.026
	(0.052)
Some college, no degree	−0.046
	(0.052)
Associate degree	−0.061
	(0.054)
Bachelor degree	−0.033
	(0.052)
Post-bachelor degree	−0.067
	(0.055)
Rep-leaning Ind	−0.004
	(0.023)

Continued on next page

Table A10 – *Continued from previous page*

	Attrited
	(1)
Weak Rep	−0.030
	(0.028)
Strong Rep	−0.048**
	(0.022)
Excuse × Age	−0.003
	(0.003)
Excuse × Age squared	0.00004
	(0.00003)
Excuse × Black	0.072
	(0.064)
Excuse × Asian	−0.027
	(0.077)
Excuse × White	0.007
	(0.051)
Excuse × Hispanic	0.019
	(0.040)
Excuse × Male	0.019
	(0.024)
Excuse × High school	0.035
	(0.068)
Excuse × Some college, no degree	0.007
	(0.068)
Excuse × Associate degree	0.051
	(0.072)
Excuse × Bachelor degree	0.062
	(0.069)
Excuse × Post-bachelor degree	0.052
	(0.073)

Continued on next page

Table A10 – *Continued from previous page*

	Attrited
	(1)
Excuse × Rep-leaning Ind	−0.004
	(0.032)
Excuse × Weak Rep	0.045
	(0.039)
Excuse × Strong Rep	0.021
	(0.031)
DV mean (no excuse)	0.151
DV mean (excuse)	0.159
Observations	3,792
R^2	0.031
Adjusted R^2	0.024

Notes: The dependent variable is an indicator that takes value 1 if the respondent attrited post-randomization. The sample is limited to respondents in the Excuse and No Excuse condition. Robust standard errors are reported.

Table A11: Experiment 2: Condition prediction confusion matrices

Panel A: *Excuse vs. No Excuse*

	Predicted Excuse	**Predicted Excuse**
True Excuse	203	254
True No Excuse	196	224
		Overall accuracy: 0.4869

Panel B: *Control vs. No Excuse*

	Predicted Excuse	**Predicted Excuse**
True Control	172	205
True No Excuse	117	302
		Overall accuracy: 0.5955

Panel C: *Control vs. Excuse*

	Predicted Excuse	**Predicted Excuse**
True Control	183	164
True Excuse	147	304
		Overall accuracy: 0.6103

Notes: Each cell reports the number of individuals who were assigned to the condition in the corresponding row and who were classified by the Support Vector Machine as belonging to the condition in the corresponding column. Each panel limits the data to the corresponding two conditions. The classifiers were trained on a 75% sample of the limited dataset; the table reports prediction results on the test set of the remaining 25% of the limited dataset. Overall accuracy is calculated as the proportion of correct predictions.

B Theoretical Results

B.1 Proof for the "no excuse" condition

B.1.1 Analysis for the receiver

In the no excuse condition, the receiver believes that the senders do not receive any signal, i.e. senders can only follow their tolerance types and the prior beliefs about the news veracity to make donation decisions. Therefore, the receiver cannot tell apart the sophisticated and gullible senders and believes that there are only two possible probabilities of a certain type of sender choosing to donate ρ_H and ρ_L and his belief updating does not reply on his own signal. Seeing different actions of senders', the receiver updates his beliefs

$$\mathbb{M}(a=1) = pr_j(\theta = H|a=1) = \frac{\pi\rho_H}{\pi\rho_H + (1-\pi)\rho_L} \tag{11}$$

$$\mathbb{M}(a=0) = pr_j(\theta = H|a=0) = \frac{\pi(1-\rho_H)}{\pi(1-\rho_H) + (1-\pi)(1-\rho_L)} \tag{12}$$

B.1.2 Analysis for the sender

Given that the senders know their own signals , and make donation decisions based on their signals. We can denote that in the Perfect Bayesian Equilibrium, the sender believes the receiver updates his type inference with

$$\widetilde{\rho}_{\theta_i} = \delta[\alpha_i\rho^T_{\theta_i}(S) + (1-\alpha_i)\rho^F_{\theta_i}(S)] + (1-\delta)[\beta\rho^T_{\theta_i}(N) + (1-\beta)\rho^F_{\theta_i}(N)] \tag{13}$$

where $\alpha_i = pr_i(s_{i'} = T|s_i, \gamma_i, \gamma_{i'} = S)$ represents the probability of another sender i' who is sophisticated and has the same tolerance type as sender i, receiving the same signal $s_{i'}$ as sender i receives given sender i's sophistication level. And $\beta_i = pr_i(s_{i'} = T|s_i, \gamma_i, \gamma_{i'} = N)$ represents the probability of another sender i' who is gullible and has the same tolerance type as sender i, receiving the same signal $s_{i'}$ as sender i receives given sender i's sophistication level.

Then for senders with different sophistication levels and different signals, α_i and β_i are

$$\begin{aligned}
(\gamma_i = S, s_i = T):&\ \alpha_i = 1, \beta_i = 1\\
(\gamma_i = S, s_i = F):&\ \alpha_i = 0, \beta_i = q\\
(\gamma_i = N, s_i = T):&\ \alpha_i = \frac{p}{p+q(1-p)}, \beta_i = \frac{p+q^2(1-p)}{p+q(1-p)}\\
(\gamma_i = N, s_i = F):&\ \alpha_i = 0, \beta_i = q
\end{aligned}$$

which in turn gives us $\widetilde{\rho}^{s_i}_\theta(S) \geq \widetilde{\rho}^{s_i}_\theta(N)$ and $\widetilde{\rho}^T_\theta(\gamma_i) \geq \widetilde{\rho}^F_\theta(\gamma_i)$.

For a sophisticated sender, when he observes $s = T$, the utility functions he has given different

actions can be written as

$$u_i(a=1) = R_i + \lambda pr_j[\pi \mathbb{M}(a=1) + (1-\pi)(1-\mathbb{M}(a=1))] \tag{14}$$

$$u_i(a=0) = -c + \lambda pr_j[\pi \mathbb{M}(a=0) + (1-\pi)(1-\mathbb{M}(a=0))] \tag{15}$$

Denote $\nabla_i^1 = pr_j[\pi \mathbb{M}(a=1) + (1-\pi)(1-\mathbb{M}(a=1))]$ and $\nabla_i^0 = pr_j[\pi \mathbb{M}(a=0) + (1-\pi)(1-\mathbb{M}(a=0))]$. And with the updated $\widetilde{\rho}_H$ and $\widetilde{\rho}_L$ in sender i's belief, we write the social rewards in sender i''s utility functions as $\lambda \widetilde{\nabla}_i^1$ and $\lambda \widetilde{\nabla}_i^0$. Then we have

$$u_i(a=1) - u_i(a=0) = R_i + c + \lambda(\widetilde{\nabla}_i^1 - \widetilde{\nabla}_i^0) \tag{16}$$

With $\lambda > \frac{R_L + c}{(2\pi - 1)\pi}$ we see that $R_L + c + \lambda(\widetilde{\nabla}_i^1 - \widetilde{\nabla}_i^0) < 0$, i.e.. $\rho_L^T(S) = 0$. Given $\rho_\theta^{s_i}(S) \geq \rho_\theta^{s_i}(N)$, $\rho_\theta^T(\gamma_i) \geq \rho_\theta^F(\gamma_i)$ and $\rho_L^{s_i}(\gamma_i) \geq \rho_H^{s_i}(\gamma_i)$, all the other types choose to donate with probability 0 as well.

B.2 Proof for the "excuse" condition

B.2.1 Analysis for the receiver

In the "excuse" condition, the receiver makes inference about the sender's signal s_i based on his own signal s_j, and subsequently, make inferences about the sender's type. Given the assumptions we have, the receiver's beliefs about the sender's signals are updated in the same way that sender i updates his belief about the sender i''s signal in the previous section, i.e.

$$\begin{aligned}
&pr_j^S(s_i^S = T | s_j^S = T) = pr_j^S(s_i^N = T | s_j^S = T) = 1 \\
&pr_j^S(s_i^S = T | s_j^S = F) = 0, \quad pr_j^S(s_i^N = T | s_j^S = F) = q2ber \\
&pr_j^N(s_i^S = T | s_j^N = T) = \frac{p}{p + q(1-p)}, \quad pr_j^N(s_i^N = T | s_j^N = T) = \frac{p + q^2(1-p)}{p + q(1-p)} \\
&pr_j^N(s_i^S = T | s_j^N = F) = 0, \quad pr_j^N(s_i^N = T | s_j^N = F) = q
\end{aligned} \tag{17}$$

When the receiver is sophisticated, and he receives a true signal about the news $s_j = T$, his

inferences about the sender's types based on the sender's donation actions should be

$$\begin{aligned}\mathbb{M}_T^S(a=1) &= pr_j(\theta=H|a=1)\\ &= \frac{[\delta\rho_H^T(S)+(1-\delta)\rho_H^T(N)]\pi}{[\delta\rho_H^T(S)+(1-\delta)\rho_H^T(N)]\pi+[\delta\rho_L^T(S)+(1-\delta)\rho_L^T(N)](1-\pi)} \qquad (18)\\ \mathbb{M}_T^S(a=0) &= pr_j(\theta=H|a=0)\\ &= \frac{[\delta(1-\rho_H^T(S))+(1-\delta)(1-\rho_H^T(N))]\pi}{[\delta(1-\rho_H^T(S))+(1-\delta)(1-\rho_H^T(N))]\pi+[\delta(1-\rho_L^T(S))+(1-\delta)(1-\rho_L^T(N))](1-\pi)} \qquad (19)\end{aligned}$$

While if he receives a false signal $s_j = F$, his inferences are updated to be

$$\begin{aligned}\mathbb{M}_F^S(a=1) =& pr_j(\theta=H|a=1)\\ =& \left\{[\delta\rho_H^F(S)+(1-\delta)(q\rho_H^T(N)+(1-q)\rho_H^F(N))]\pi\right\}\Big/\\ & \left\{[\delta\rho_H^F(S)+(1-\delta)(q\rho_H^T(N)+(1-q)\rho_H^F(N))]\pi+\right.\\ & \left.[\delta\rho_L^F(S)+(1-\delta)(q\rho_L^T(N)+(1-q)\rho_L^F(N))](1-\pi)\right\} \qquad (20)\\ \mathbb{M}_F^S(a=0) =& pr_j(\theta=H|a=0)\\ =& \left\{[\delta(1-\rho_H^F(S))+(1-\delta)(q(1-\rho_H^T(N))+(1-q)(1-\rho_H^F(N)))]\pi\right\}\Big/\\ & \left\{[\delta(1-\rho_H^F(S))+(1-\delta)(q(1-\rho_H^T(N))+(1-q)(1-\rho_H^F(N)))]\pi\right.\\ & \left.+[\delta(1-\rho_L^F(S))+(1-\delta)(q(1-\rho_L^T(N))+(1-q)(1-\rho_L^F(N)))](1-\pi)\right\} \qquad (21)\end{aligned}$$

And when the receiver is gullible, and he receives a true signal about the news $s_j = T$, he would first update his own beliefs about the veracity of the news and then make inferences about

the sender's types:

$$\begin{aligned}
\mathbb{M}_T^N(a=1) =& pr_j(\theta=H|a=1)\\
=&\Bigg\{\Big[\delta(\frac{p}{p+(1-p)q}\rho_H^T(S)+\frac{(1-p)q}{p+(1-p)q}\rho_H^F(S))+(1-\delta)(\frac{p+q^2(1-p)}{p+(1-p)q}\rho_H^T(N)+\frac{(1-p)q(1-q)}{p+(1-p)q}\rho_H^F(N))\Big]\pi\Bigg\}\Bigg/\\
&\Bigg\{\Big[\delta(\frac{p}{p+(1-p)q}\rho_H^T(S)+\frac{(1-p)q}{p+(1-p)q}\rho_H^F(S))+(1-\delta)(\frac{p+q^2(1-p)}{p+(1-p)q}\rho_H^T(N)+\frac{(1-p)q(1-q)}{p+(1-p)q}\rho_H^F(N))\Big]\pi+\\
&\Big[\delta(\frac{p}{p+(1-p)q}\rho_L^T(S)+\frac{(1-p)q}{p+(1-p)q}\rho_L^F(S))+(1-\delta)(\frac{p+q^2(1-p)}{p+(1-p)q}\rho_L^T(N)+\frac{(1-p)q(1-q)}{p+(1-p)q}\rho_L^F(N))\Big](1-\pi)\Bigg\} \quad (22)\\
\mathbb{M}_T^N(a=0) =& pr_j(\theta=H|a=1)\\
=&\Bigg\{\Big[\delta(\frac{p}{p+(1-p)q}(1-\rho_H^T(S))+\frac{(1-p)q}{p+(1-p)q}(1-\rho_H^F(S)))+(1-\delta)(\frac{p+q^2(1-p)}{p+(1-p)q}(1-\rho_H^T(N))+\frac{(1-p)q(1-q)}{p+(1-p)q}(1-\rho_H^F(N)))\Big]\pi\Bigg\}\Bigg/\\
&\Bigg\{\Big[\delta(\frac{p}{p+(1-p)q}(1-\rho_H^T(S))+\frac{(1-p)q}{p+(1-p)q}(1-\rho_H^F(S)))+(1-\delta)(\frac{p+q^2(1-p)}{p+(1-p)q}(1-\rho_H^T(N))+\frac{(1-p)q(1-q)}{p+(1-p)q}(1-\rho_H^F(N)))\Big]\pi+\\
&\Big[\delta(\frac{p}{p+(1-p)q}(1-\rho_L^T(S))+\frac{(1-p)q}{p+(1-p)q}(1-\rho_L^F(S)))+(1-\delta)(\frac{p+q^2(1-p)}{p+(1-p)q}(1-\rho_L^T(N))+\frac{(1-p)q(1-q)}{p+(1-p)q}(1-\rho_L^F(N)))\Big](1-\pi)\Bigg\}
\end{aligned} \quad (23)$$

While if he receives a false signal $s_j = F$, the inferences the receiver make about the sender's signals should are the same as when he is a sophisticated agent and receives a false signal. Thus we denote $\mathbb{M}_F(a=1) = \mathbb{M}_F^N(a=1) = \mathbb{M}_F^S(a=1)$ and $\mathbb{M}_F(a=0) = \mathbb{M}_F^N(a=0) = \mathbb{M}_F^S(a=0)$.

B.2.2 Analysis for the sender

When the sender receives some signal s_i, the inference he makes about the receiver's signal follows the same process of the receiver making inferences about the sender's signal, which was analyzed in the previous section. We start by assuming that the sender is sophisticated and receives a signal $s_i = T$. Then his infers that the news being true happens with probability 1 and consequently, the receiver receives a true signal $s_j = T$ with probability 1. The utility function of his can be written as

$$\begin{aligned}
u_i =&(R_{\theta_i} - c\mathbf{Pr}(n=F|s_i))a - c\mathbf{Pr}(n=T|s_i)(1-a) + \lambda\Big[\pi pr_j(\theta_i=H|a) + (1-\pi)pr_j(\theta_i=L|a)\Big]\\
=&R_{\theta_i}a - c(1-a) + \lambda\Big[\pi pr_j(\theta_i=H|a) + (1-\pi)pr_j(\theta_i=L|a)\Big]\\
=&R_{\theta_i}a - c(1-a) + \lambda\nabla_i
\end{aligned} \quad (24)$$

Now the social reward term $\lambda\nabla_i$ depends on a with s_2 fixed. Then regardless of the sender's tolerance type, the social reward he receives by choosing $a=1$ and $a=2$ are

$$\lambda\nabla_T^1 = \lambda\Big[\pi\Big(\delta\mathbb{M}_T^S(a=1) + (1-\delta)\mathbb{M}_T^N(a=1)\Big) + (1-\pi)\Big(\delta(1-\mathbb{M}_T^S(a=1)) + (1-\delta)(1-\mathbb{M}_T^N(a=1))\Big)\Big] \quad (25)$$

$$\lambda\nabla_T^0 = \lambda\Big[\pi\Big(\delta\mathbb{M}_T^S(a=0) + (1-\delta)\mathbb{M}_T^N(a=0)\Big) + (1-\pi)\Big(\delta(1-\mathbb{M}_T^S(a=0)) + (1-\delta)(1-\mathbb{M}_T^N(a=0))\Big)\Big] \quad (26)$$

The utilities the sender can obtain with different donation decisions are

$$u_i(a=1) = R_\theta + \lambda \nabla_T^1 \tag{27}$$

$$u_i(a=0) = -c + \lambda \nabla_T^0 \tag{28}$$

Given that $\rho_H^T(S) \leq \rho_L^T(S)$, we first consider the case where the equality holds. Then there are two possibilities $\rho_H^T(S) = \rho_L^T(S) = 1$ and $\rho_H^T(S) = \rho_L^T(S) = 0$. For the first equation to be true, we must have $R_H + c + \lambda(\nabla_T^1 - \nabla_T^0) > 0$ and it being true implies that $\rho_H^T(N) = \rho_L^T(N) = 1$. Then we update $\mathbb{M}_T^S(a=1) = \pi$, $\mathbb{M}_T^S(a=0) = 1$, and

$$\begin{aligned}\mathbb{M}_T^N(a=1) = &\left\{ \Big[\delta(p + (1-p)q\rho_H^F(S)) + (1-\delta)((p + q^2(1-p)) + (1-p)q(1-q)\rho_H^F(N))\Big]\pi \right\} \Big/ \\ &\left\{ \Big[\delta(p + (1-p)q\rho_H^F(S)) + (1-\delta)((p + q^2(1-p)) + (1-p)q(1-q)\rho_H^F(N))\Big]\pi + \right. \\ &\left. \Big[\delta(p + (1-p)q\rho_L^F(S)) + (1-\delta)((p + q^2(1-p)) + (1-p)q(1-q)\rho_L^F(N))\Big](1-\pi) \right\}\end{aligned} \tag{29}$$

$$\begin{aligned}\mathbb{M}_T^N(a=0) = &\left\{ \Big[\delta(1-p)q(1-\rho_H^F(S)) + (1-\delta)(1-p)q(1-q)(1-\rho_H^F(N))\Big]\pi \right\} \Big/ \\ &\left\{ \Big[\delta(1-p)q(1-\rho_H^F(S)) + (1-\delta)(1-p)q(1-q)(1-\rho_H^F(N))\Big]\pi + \right. \\ &\left. \Big[\delta(1-p)q(1-\rho_L^F(S)) + (1-\delta)(1-p)q(1-q)(1-\rho_L^F(N))\Big](1-\pi) \right\}\end{aligned} \tag{30}$$

However, since $\lambda > \frac{R_L + c}{(2\pi - 1)\pi}$ and $R_L > c$, we have $\lambda > 2c$. And under this condition, there exists $\pi \in (\frac{1}{2}, 1]$ that makes $R_H + c + \lambda(\nabla_T^1 - \nabla_T^0) \leq 0$ regardless of the values of $\rho_H^F(S)$ and $\rho_H^F(N)$. So the equilibrium dissolves.

Now consider the second possibility $\rho_H^T(S) = \rho_L^T(S) = 0$. Similar to before, there is no equilibrium where $\rho_H^T(S) = \rho_L^T(S) = \rho_H^T(N) = \rho_L^T(N) = 0$ as it is not incentive compatible, so we must have $\rho_H^T(S) < \rho_H^T(N)$ and $\rho_L^T(S) < \rho_L^T(N)$. We look at the corner solutions first. Suppose that $\rho_H^T(N) = 1$. This implies that $\rho_L^T(N) = 1$ since the intolerant sender can be no worse off than the tolerant sender by choosing $a = 1$. Then the receiver's inferences can be updated as

$\mathbb{M}_T^S(a=1)=\pi$, $\mathbb{M}_T^S(a=0)=1$, and

$$\begin{aligned}\mathbb{M}_T^N(a=1)=&\left\{\left[\delta(1-p)q\rho_H^F(S)+(1-\delta)((p+q^2(1-p))+(1-p)q(1-q)\rho_H^F(N))\right]\pi\right\}\Big/\\&\left\{\left[\delta(1-p)q\rho_H^F(S)+(1-\delta)((p+q^2(1-p))+(1-p)q(1-q)\rho_H^F(N))\right]\pi+\right.\\&\left.\left[\delta(1-p)q\rho_L^F(S)+(1-\delta)((p+q^2(1-p))+(1-p)q(1-q)\rho_L^F(N))\right](1-\pi)\right\}\end{aligned}\tag{31}$$

$$\begin{aligned}\mathbb{M}_T^N(a=0)=&\left\{\left[\delta(p+(1-p)q(1-\rho_H^F(S)))+(1-\delta)(1-p)q(1-q)(1-\rho_H^F(N))\right]\pi\right\}\Big/\\&\left\{\left[\delta(p+(1-p)q(1-\rho_H^F(S)))+(1-\delta)(1-p)q(1-q)(1-\rho_H^F(N))\right]\pi+\right.\\&\left.\left[\delta(p+(1-p)q(1-\rho_L^F(S)))+(1-\delta)(1-p)q(1-q)(1-\rho_L^F(N))\right](1-\pi)\right\}\end{aligned}\tag{32}$$

For the equilibrium to hold, we need $c+\lambda(\nabla_T^1-\nabla_T^0)>0$ and $R+(1-\frac{q(1-p)}{p+q(1-p)})c+\lambda\Big[\frac{\delta p+(1-\delta)(p+q^2(1-p))}{p+q(1-p)}(\nabla_T^1-\nabla_T^0)+\frac{\delta q(1-p)+(1-\delta)q(1-q)(1-p)}{p+q(1-p)}(\nabla_F^1-\nabla_F^0)\Big]\leq 0$. And here

$$\lambda(\nabla_F^1-\nabla_F^0)=\lambda\Big[\pi(\mathbb{M}_F(a=1)-\mathbb{M}_F(a=0))+(1-\pi)((1-\mathbb{M}_F(a=1))-(1-\mathbb{M}_F(a=0)))\Big]\tag{33}$$

with the updated beliefs

$$\begin{aligned}\mathbb{M}_F(a=1)=&\left\{[\delta\rho_H^F(S)+(1-\delta)(q+(1-q)\rho_H^F(N))]\pi\right\}\Big/\\&\left\{[\delta\rho_H^F(S)+(1-\delta)(q+(1-q)\rho_H^F(N))]\pi+[\delta\rho_L^F(S)+(1-\delta)(q+(1-q)\rho_L^F(N))](1-\pi)\right\}\end{aligned}\tag{34}$$

$$\begin{aligned}\mathbb{M}_F(a=0)=&\left\{[\delta(1-\rho_H^F(S))+(1-q)(1-\rho_H^F(N))]\pi\right\}\Big/\\&\left\{[\delta(1-\rho_H^F(S))+(1-q)(1-\rho_H^F(N))]\pi+[\delta(1-\rho_L^F(S))+(1-q)(1-\rho_L^F(N))](1-\pi)\right\}\end{aligned}\tag{35}$$

Taking the values in, we see that the equilibrium dissolves and $\nabla_T^1 - \nabla_T^0$ decreases with $\rho_H^T(N)$, so any $\rho_H^T(N) < 1$ dissolves the equilibrium where $\rho_H^T(S) = \rho_L^T(S) = 0$. Therefore, the second equality condition does not yield any equilibrium.

Now we consider the inequality condition $\rho_H^T(S) < \rho_L^T(S)$. For simplicity, we rule out the possibilities of the interior solutions, thus we must have $\rho_H^T(S) = 0$ and $\rho_L^T(S) = 1$. This further implies that $\rho_L^T(N) = 1$.

We start by assuming that $\rho_H^T(N) = 0$. This updates us with $\mathbb{M}_T^S(a = 1) = 0$, $\mathbb{M}_T^S(a = 0) = 1$, and

$$\begin{aligned}\mathbb{M}_T^N(a=1) =& \Big\{[\delta(1-p)q\rho_H^F(S) + (1-\delta)(1-p)q(1-q)\rho_H^F(N)]\pi\Big\} \Big/ \\ & \Big\{[\delta(1-p)q\rho_H^F(S) + (1-\delta)(1-p)q(1-q)\rho_H^F(N)]\pi + \\ & [\delta(p+(1-p)q\rho_L^F(S)) + (1-\delta)((p+q^2(1-p)) + (1-p)q(1-q)\rho_L^F(N))](1-\pi)\Big\} \end{aligned} \tag{36}$$

$$\begin{aligned}\mathbb{M}_T^N(a=0) =& \Big\{[\delta(p+(1-p)q(1-\rho_H^F(S))) + (1-\delta)((p+q^2(1-p)) + (1-p)q(1-q)\rho_H^F(N))]\pi\Big\} \Big/ \\ & \Big\{[\delta(p+(1-p)q(1-\rho_H^F(S))) + (1-\delta)((p+q^2(1-p)) + (1-p)q(1-q)\rho_H^F(N))]\pi + \\ & [\delta(1-p)q(1-\rho_L^F(S)) + (1-\delta)(1-p)q(1-q)(1-\rho_L^F(N)))](1-\pi)\Big\} \end{aligned} \tag{37}$$

$$\begin{aligned}\mathbb{M}_F^S(a=1) =& \Big\{[\delta\rho_H^F(S) + (1-\delta)(1-q)\rho_H^F(N)]\pi\Big\} \Big/ \\ & \Big\{[\delta\rho_H^F(S) + (1-\delta)(q\rho_H^T(N) + (1-q)\rho_H^F(N))]\pi + [\delta\rho_L^F(S) + (1-\delta)(1-q)\rho_L^F(N)](1-\pi)\Big\} \end{aligned} \tag{38}$$

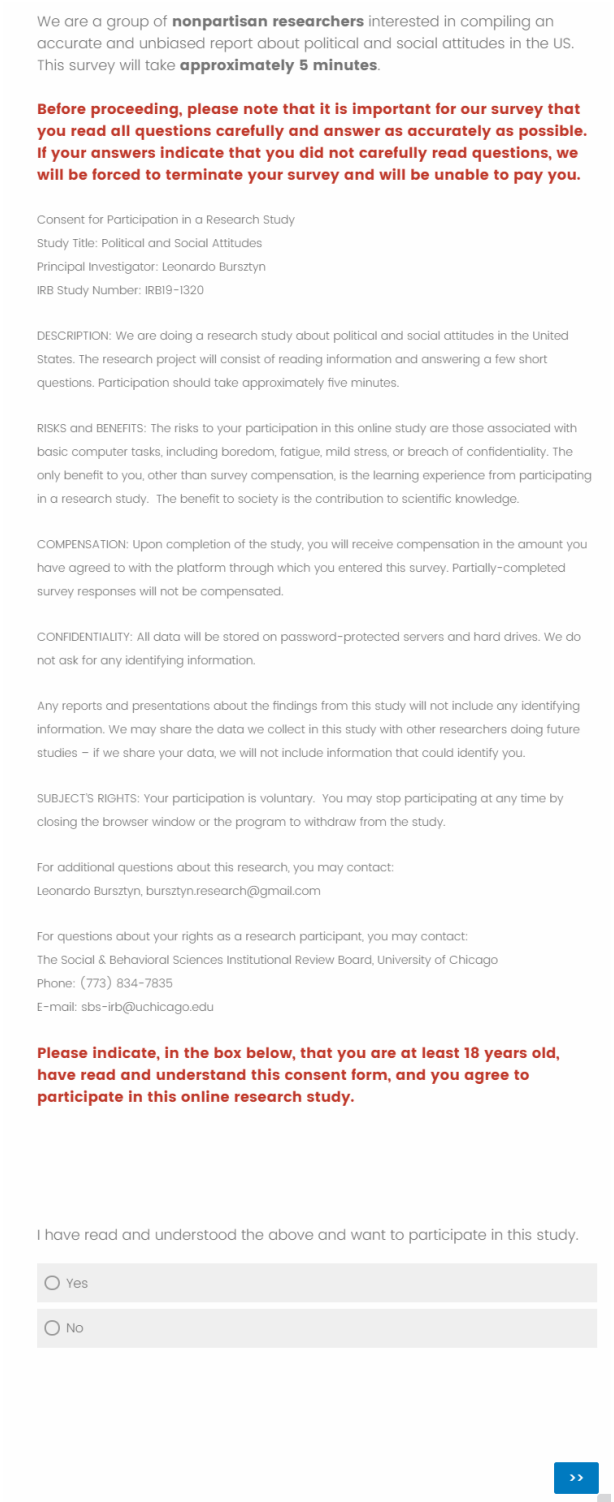
$$\begin{aligned}\mathbb{M}_F^S(a=0) =& \Big\{[\delta(1-\rho_H^F(S)) + (1-\delta)(q+(1-q)(1-\rho_H^F(N)))]\pi\Big\} \Big/ \\ & \Big\{[\delta(1-\rho_H^F(S)) + (1-\delta)(q+(1-q)(1-\rho_H^F(N)))]\pi + \\ & [\delta(1-\rho_L^F(S)) + (1-\delta)(1-q)(1-\rho_L^F(N))](1-\pi)\Big\} \end{aligned} \tag{39}$$

With the updated beliefs and $(2\pi - 1)\pi\lambda > R_L + c$, $R + c + \lambda(\nabla_T^1 - \nabla_T^0) < 0$, contradicting

$\rho_L^T(S) = 1.$

C Survey instruments: Experiment 1

C.1 Consent and pre-treatment questions



We are a group of **nonpartisan researchers** interested in compiling an accurate and unbiased report about political and social attitudes in the US. This survey will take **approximately 5 minutes**.

Before proceeding, please note that it is important for our survey that you read all questions carefully and answer as accurately as possible. If your answers indicate that you did not carefully read questions, we will be forced to terminate your survey and will be unable to pay you.

Consent for Participation in a Research Study
Study Title: Political and Social Attitudes
Principal Investigator: Leonardo Bursztyn
IRB Study Number: IRB19-1320

DESCRIPTION: We are doing a research study about political and social attitudes in the United States. The research project will consist of reading information and answering a few short questions. Participation should take approximately five minutes.

RISKS and BENEFITS: The risks to your participation in this online study are those associated with basic computer tasks, including boredom, fatigue, mild stress, or breach of confidentiality. The only benefit to you, other than survey compensation, is the learning experience from participating in a research study. The benefit to society is the contribution to scientific knowledge.

COMPENSATION: Upon completion of the study, you will receive compensation in the amount you have agreed to with the platform through which you entered this survey. Partially-completed survey responses will not be compensated.

CONFIDENTIALITY: All data will be stored on password-protected servers and hard drives. We do not ask for any identifying information.

Any reports and presentations about the findings from this study will not include any identifying information. We may share the data we collect in this study with other researchers doing future studies – if we share your data, we will not include information that could identify you.

SUBJECT'S RIGHTS: Your participation is voluntary. You may stop participating at any time by closing the browser window or the program to withdraw from the study.

For additional questions about this research, you may contact:
Leonardo Bursztyn, bursztyn.research@gmail.com

For questions about your rights as a research participant, you may contact:
The Social & Behavioral Sciences Institutional Review Board, University of Chicago
Phone: (773) 834-7835
E-mail: sbs-irb@uchicago.edu

Please indicate, in the box below, that you are at least 18 years old, have read and understand this consent form, and you agree to participate in this online research study.

I have read and understood the above and want to participate in this study.

○ Yes

○ No

>>

The next question is about the following problem. In questionnaires like ours, sometimes there are participants who do not carefully read the questions and just quickly click through the survey. This means that there are a lot of random answers which compromise the results of research studies. **To show that you read our questions carefully, please choose both "Extremely interested" and "Not at all interested" as your answer in the next question.** How interested are you in sports?

- ☐ Extremely interested
- ☐ Very interested
- ☐ A little bit interested
- ☐ Almost not interested
- ☐ Not at all interested

>>

Are you Spanish, Hispanic, or Latino or none of these?

- ○ Yes
- ○ None of these

What is your year of birth?

What is your sex?

- ○ Male
- ○ Female

In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?

- ○ Republican
- ○ Democrat
- ○ Independent

>>

Do you strongly support or weakly support the Democratic Party?

- ◯ Strongly support
- ◯ Weakly support

What is the highest level of school you have completed or the highest degree you have received?

- ◯ Less than high school degree
- ◯ High school graduate (high school diploma or equivalent including GED)
- ◯ Some college but no degree
- ◯ Associate degree in college (2-year)
- ◯ Bachelor's degree in college (4-year)
- ◯ Master's degree
- ◯ Doctoral degree
- ◯ Professional degree (JD, MD)

Which of the following best describes your race or ethnicity?

- ◯ African American/Black
- ◯ Asian/Asian American
- ◯ Caucasian/White
- ◯ Native American, Inuit or Aleut
- ◯ Native Hawaiian/Pacific Islander
- ◯ Other

>>

Do Illegal Immigrants Commit More Crime? Evidence from Dr. Lott's Study

Dr. John R. Lott, an economist formerly employed at top institutions such as Yale University and the University of Chicago, carried out a study on the relationship between illegal immigration and crime using new high-quality data. **The study has not yet been published in an academic journal, but we obtained an early version and summarize the results below.**

Dr. Lott's study uses **detailed data spanning over two decades from the Arizona state prison systems**. He finds that **"undocumented immigrants are at least 142% more likely** to be convicted of a crime than other Arizonans. They also tend to commit **more serious crimes**," such as murder and rape.

Moreover, "young undocumented immigrants commit crime at twice the rate of young U.S. citizens." He calculates that "if undocumented immigrants committed crime nationally as they do in Arizona, in 2016 they would have been responsible for over **1,000 more murders, 5,200 rapes, 8,900 robberies, 25,300 aggravated assaults, and 26,900 burglaries**. The vast majority of the illegal immigrants Lott studies are from Mexico.

However, a number of sources (including a researcher affiliated with the Cato Institute, a non-partisan think tank) have recently challenged some of the study's methods, claiming that errors in analysis invalidate its results. Dr. Lott has responded to this critique with a defense of the study's methods, but the issue remains unresolved.

>>

C.2 No excuse condition

We conducted a survey about political and social attitudes in the United States earlier this year. You have been matched with one of the respondents from that survey.

We gave your matched respondent the opportunity to authorize a $1 donation to **Fund the Wall**, a nonprofit organization that seeks to reduce illegal immigration into the United States by helping to fund and construct the US-Mexico border wall. Your matched respondent was told that their donation decision would be posted on our website. The decision on whether to authorize the donation did not have any financial consequences for your matched respondent.

Some respondents were assigned a longer version of the survey and learned about Dr. Lott's study before they decided whether or not to donate. Other respondents were assigned a shorter version of the study and **were not informed** about Dr. Lott's study before they decided whether or not to donate.

Information about your matched respondent

- Your matched respondent **was not informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent decided to authorize the $1 donation to Fund the Wall

>>

Why do you think your matched respondent chose to donate to Fund the Wall?

Reminder: Information about your matched respondent

- Your matched respondent **was not informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent decided to authorize the $1 donation to Fund the Wall

>>

After your matched respondent made their donation decision, they completed the **The Gullibility Scale**, a short questionnaire which measures **how easily people are manipulated by evidence from untrustworthy sources**.

On the next page, we will ask you to guess how your matched respondent scored on this scale. If you guess the correct option, you will be entered into a lottery for a $50 Amazon gift card.

>>

The Gullibility Scale

We administered **The Gullibility Scale**, a short questionnaire which measures **how easily people are manipulated by evidence from untrustworthy sources**, to your matched respondent.

The test is scored from 0 to 100, where 0 means "least gullible" and 100 means "most gullible". Thus, a higher score indicates that your matched respondent is more gullible.

Reminder: Information about your matched respondent

- Your matched respondent **was not informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent decided to authorize the $1 donation to Fund the Wall

If you had to guess, how do you think your **matched respondent** scored on **The Gullibility Scale**?

- ◯ Score between 0 and 10 (**Not at all gullible**)
- ◯ Score between 10 and 20
- ◯ Score between 20 and 30
- ◯ Score between 30 and 40
- ◯ Score between 40 and 50
- ◯ Score between 50 and 60
- ◯ Score between 60 and 70
- ◯ Score between 70 and 80
- ◯ Score between 80 and 90
- ◯ Score between 90 and 100 (**Extremely gullible**)

>>

After your matched respondent made their donation decision, they completed the **Foreign Culture Tolerance Scale**, a short questionnaire which measures **tolerance toward foreign values and traditions**.

On the next page, we will ask you to guess how your matched respondent scored on this scale. If you guess the correct option, you will be entered into a lottery for a $50 Amazon gift card.

>>

The Foreign Culture Tolerance Scale

We administered the **Foreign Culture Tolerance Scale**, a short questionnaire which measures tolerance toward **foreign values and traditions**, to your matched respondent.

The test is scored from 0 to 100, where 0 means "least tolerant" and 100 means "most tolerant". Thus, a **higher score indicates that your matched respondent is more tolerant toward foreign values and traditions.**

Reminder: Information about your matched respondent

- Your matched respondent **was not informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent decided to authorize the $1 donation to Fund the Wall

If you had to guess, how do you think your **matched respondent** scored on the **Foreign Culture Tolerance Scale**?

- ◯ Score between 0 and 10 (**Not at all tolerant**)
- ◯ Score between 10 and 20
- ◯ Score between 20 and 30
- ◯ Score between 30 and 40
- ◯ Score between 40 and 50
- ◯ Score between 50 and 60
- ◯ Score between 60 and 70
- ◯ Score between 70 and 80
- ◯ Score between 80 and 90
- ◯ Score between 90 and 100 (**Extremely tolerant**)

>>

C.3 Excuse condition

We conducted a survey about political and social attitudes in the United States earlier this year. You have been matched with one of the respondents from that survey.

We gave your matched respondent the opportunity to authorize a $1 donation to **Fund the Wall**, a nonprofit organization that seeks to reduce illegal immigration into the United States by helping to fund and construct the US-Mexico border wall. Your matched respondent was told that their donation decision would be posted on our website. The decision on whether to authorize the donation did not have any financial consequences for your matched respondent.

Some respondents were assigned a longer version of the survey and learned about Dr. Lott's study before they decided whether or not to donate. Other respondents were assigned a shorter version of the study and **were not informed** about Dr. Lott's study before they decided whether or not to donate.

Information about your matched respondent

- Your matched respondent **was informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent then decided to authorize the $1 donation to Fund the Wall

>>

Why do you think your matched respondent chose to donate to Fund the Wall?

Reminder: Information about your matched respondent

- Your matched respondent **was informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent then decided to authorize the $1 donation to Fund the Wall

>>

After your matched respondent made their donation decision, they completed the **The Gullibility Scale**, a short questionnaire which measures **how easily people are manipulated by evidence from untrustworthy sources**.

On the next page, we will ask you to guess how your matched respondent scored on this scale. If you guess the correct option, you will be entered into a lottery for a $50 Amazon gift card.

>>

The Gullibility Scale

We administered **The Gullibility Scale**, a short questionnaire which measures **how easily people are manipulated by evidence from untrustworthy sources**, to your matched respondent.

The test is scored from 0 to 100, where 0 means "least gullible" and 100 means "most gullible". Thus, a higher score indicates that your matched respondent is more gullible.

Reminder: Information about your matched respondent

- Your matched respondent **was informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent then decided to authorize the $1 donation to Fund the Wall

If you had to guess, how do you think your **matched respondent** scored on **The Gullibility Scale**?

- ◯ Score between 0 and 10 (**Not at all gullible**)
- ◯ Score between 10 and 20
- ◯ Score between 20 and 30
- ◯ Score between 30 and 40
- ◯ Score between 40 and 50
- ◯ Score between 50 and 60
- ◯ Score between 60 and 70
- ◯ Score between 70 and 80
- ◯ Score between 80 and 90
- ◯ Score between 90 and 100 (**Extremely gullible**)

>>

After your matched respondent made their donation decision, they completed the **Foreign Culture Tolerance Scale**, a short questionnaire which measures **tolerance toward foreign values and traditions**.

On the next page, we will ask you to guess how your matched respondent scored on this scale. If you guess the correct option, you will be entered into a lottery for a $50 Amazon gift card.

>>

The Foreign Culture Tolerance Scale

We administered the **Foreign Culture Tolerance Scale**, a short questionnaire which measures tolerance toward **foreign values and traditions**, to your matched respondent.

The test is scored from 0 to 100, where 0 means "least tolerant" and 100 means "most tolerant". Thus, a **higher score indicates that your matched respondent is more tolerant toward foreign values and traditions.**

Reminder: Information about your matched respondent

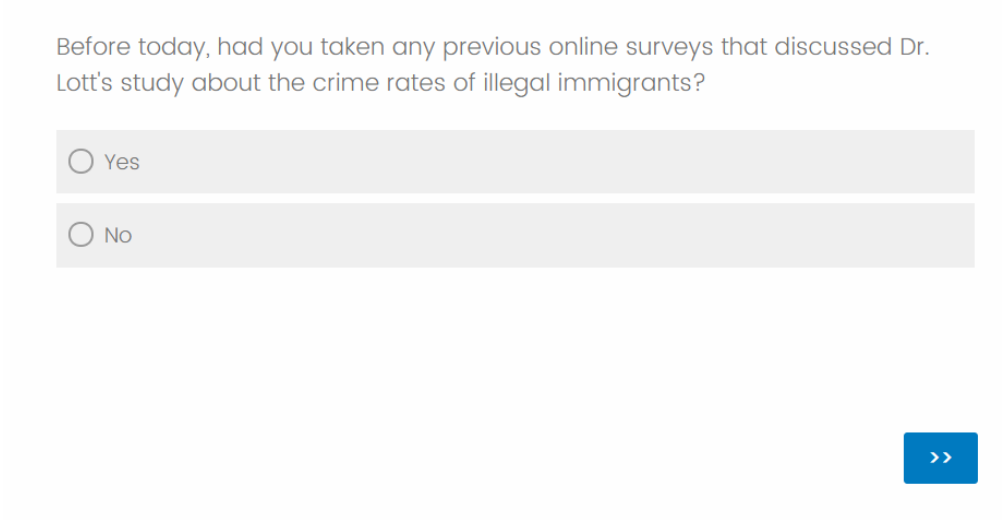
- Your matched respondent **was informed about Dr. Lott's study**, which finds that illegal immigrants commit more crimes than US citizens
- Your matched respondent then decided to authorize the $1 donation to Fund the Wall

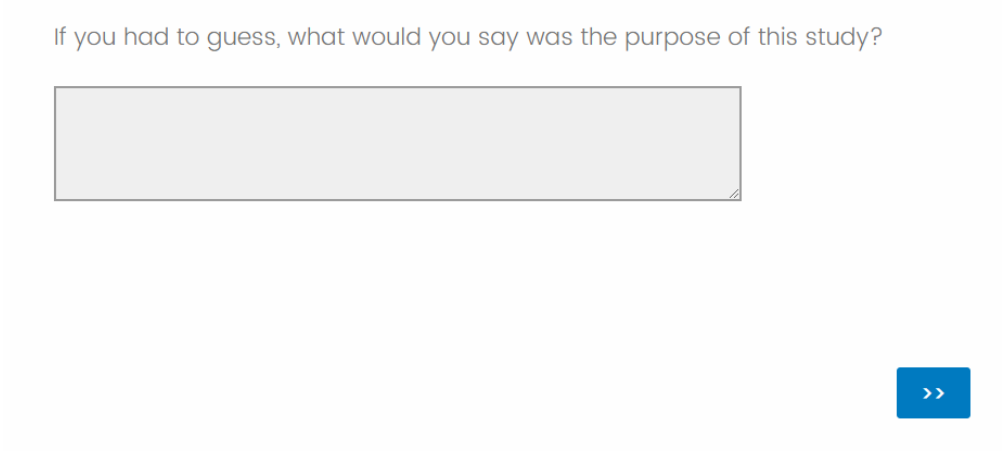
If you had to guess, how do you think your **matched respondent** scored on the **Foreign Culture Tolerance Scale**?

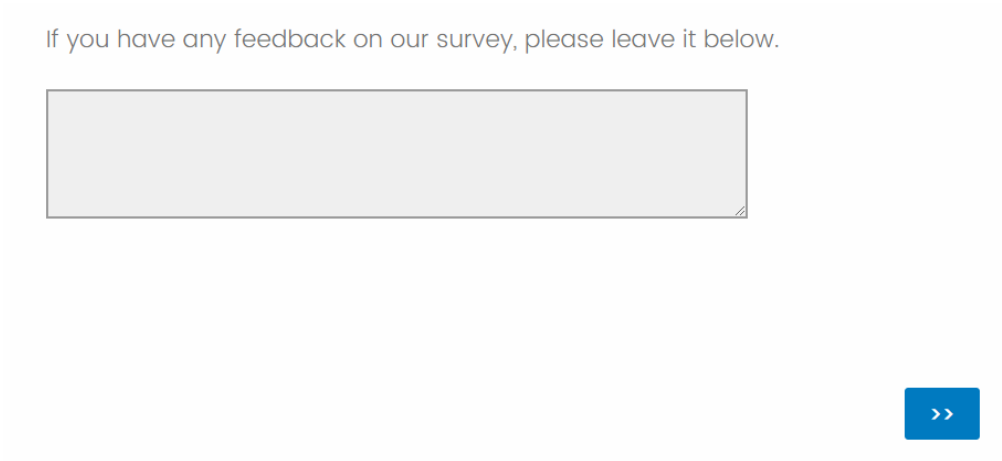
- ◯ Score between 0 and 10 (**Not at all tolerant**)
- ◯ Score between 10 and 20
- ◯ Score between 20 and 30
- ◯ Score between 30 and 40
- ◯ Score between 40 and 50
- ◯ Score between 50 and 60
- ◯ Score between 60 and 70
- ◯ Score between 70 and 80
- ◯ Score between 80 and 90
- ◯ Score between 90 and 100 (**Extremely tolerant**)

>>

C.4 Post-treatment questions and debrief







Thanks for completing all our questions!

Study Title: Political and Social Attitudes
Principal Investigator: Leonardo Bursztyn
IRB Study Number: IRB19-1320

Debrief about crime rates among immigrants in the survey

We earlier provided you with truthful information about Dr Lott's study. As we mentioned earlier in the survey, Dr. Lott's study has been challenged by a number of sources for inaccuracies in data analysis. While his methods have not been entirely debunked, there remains a great deal of controversy. Due to these problems, it is unclear whether Dr. Lott's results will be published in a reputable academic journal. We did not expand upon this controversy during the study, but if you wish to read more, we suggest this analysis by a researcher at the Cato Institute, a nonpartisan think-tank.

Immigration and crime refers to perceived or actual relationships between crime and immigration. The academic literature provides mixed findings for the relationship between immigration and crime worldwide, but finds for the United States that immigration either has no impact on the crime rate or that it reduces the crime rate. A meta-analysis of 51 studies from 1994–2014 on the relationship between immigration and crime in the United States found that overall immigration reduces crime, but the relationship is very weak. Research suggests that people tend to overestimate the relationship between immigration and criminality,and that the media tends to erroneously depict immigrants as particularly crime-prone.

The relevant meta-analysis we are referring to is the following article.
Ousey, Graham C.; Kubrin, Charis E. (2018). "Immigration and Crime: Assessing a Contentious Issue". *Annual Review of Criminology*. **1** (1): null.

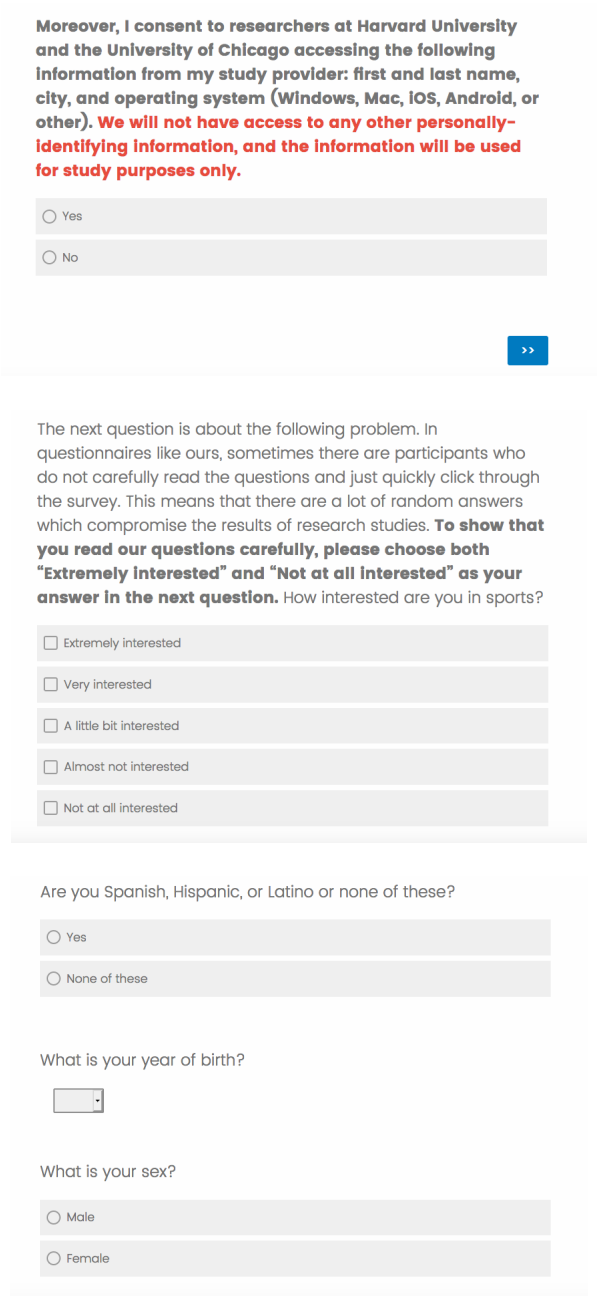
To get more information on this meta-analysis, clock on the link below:
doi:10.1146/annurev-criminol-032317-092026.

Contacts & Questions:
If you have questions or concerns about the study, you can contact the researchers at bursztyn.research@gmail.com.

Final Report: If you would like to receive a report of this study (or a summary of the findings) when it is completed, contact the researcher at the email address or phone number above.

>>

D Survey instruments: Experiment 2



In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?

- ○ Republican
- ○ Democrat
- ○ Independent

>>

Do you strongly support or weakly support the Republican Party?

- ○ Strongly support
- ○ Weakly support

>>

What is the highest level of school you have completed or the highest degree you have received?

- ○ Less than high school degree
- ○ High school graduate (high school diploma or equivalent including GED)
- ○ Some college but no degree
- ○ Associate degree in college (2-year)
- ○ Bachelor's degree in college (4-year)
- ○ Master's degree
- ○ Doctoral degree
- ○ Professional degree (JD, MD)

Which of the following best describes your race or ethnicity?

- ○ African American/Black
- ○ Asian/Asian American
- ○ Caucasian/White
- ○ Native American, Inuit or Aleut
- ○ Native Hawaiian/Pacific Islander
- ○ Other

>>

Please read the following information carefully, as it is important for the rest of the survey.

>>

Studying the crime rates of illegal immigrants is difficult. As the independent fact-checking organization PolitiFact reported in 2016, "The challenge in finding concrete numbers is due to a shortfall of data. There is no national database or study tracking how many people have been killed by undocumented immigrants or the nationality of the victims."

Therefore, most previous research has used flawed data.

Information about Lott Study: Excuse and no excuse condition

Do Illegal Immigrants Commit More Crime? Evidence from Dr. Lott's Study

Dr. John R. Lott, an economist formerly employed at top institutions such as Yale University and the University of Chicago, carried out a study on the relationship between illegal immigration and crime using new high-quality data. **The study has not yet been published in an academic journal, but we obtained an early version and summarize the results below.**

Dr. Lott's study uses **detailed data spanning over two decades from the Arizona state prison systems**. He finds that **"undocumented immigrants are at least 142% more likely** to be convicted of a crime than other Arizonans. They also tend to commit **more serious crimes**," such as murder and rape.

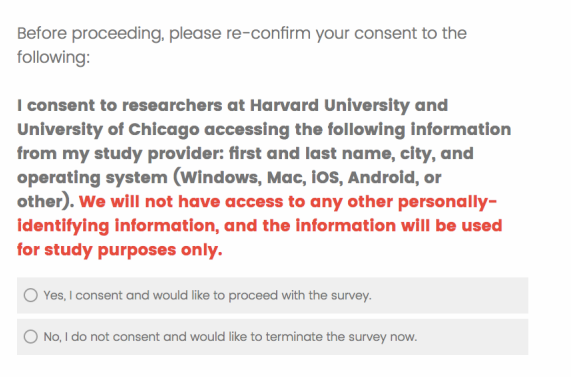
Moreover, "young undocumented immigrants commit crime at twice the rate of young U.S. citizens." He calculates that "if undocumented immigrants committed crime nationally as they do in Arizona, in 2016 they would have been responsible for over **1,000 more murders, 5,200 rapes, 8,900 robberies, 25,300 aggravated assaults, and 26,900 burglaries**. The vast majority of the illegal immigrants Lott studies are from Mexico.

However, a number of sources (including a researcher affiliated with the Cato Institute, a non-partisan think tank) have recently challenged some of the study's methods, claiming that errors in analysis invalidate its results. Dr. Lott has responded to this critique with a defense of the study's methods, but the issue remains unresolved.

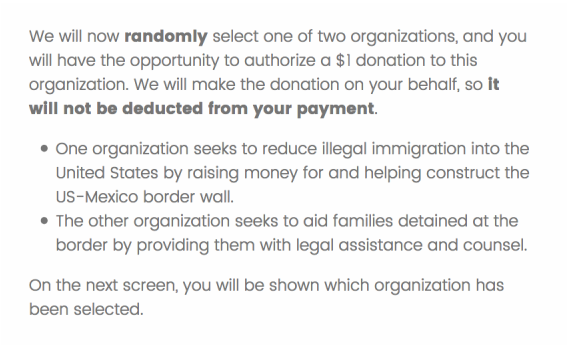
Do you think that Dr. Lott's study will be widely discussed in the media when it is published in an academic journal?

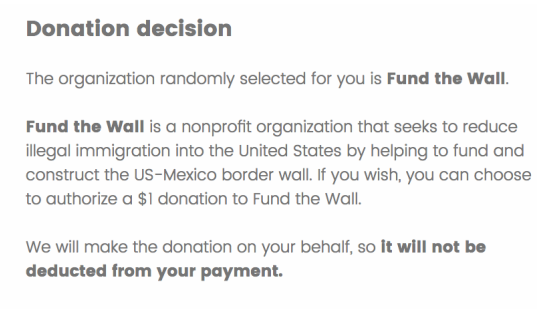
- ○ Yes
- ○ No

Reconsent

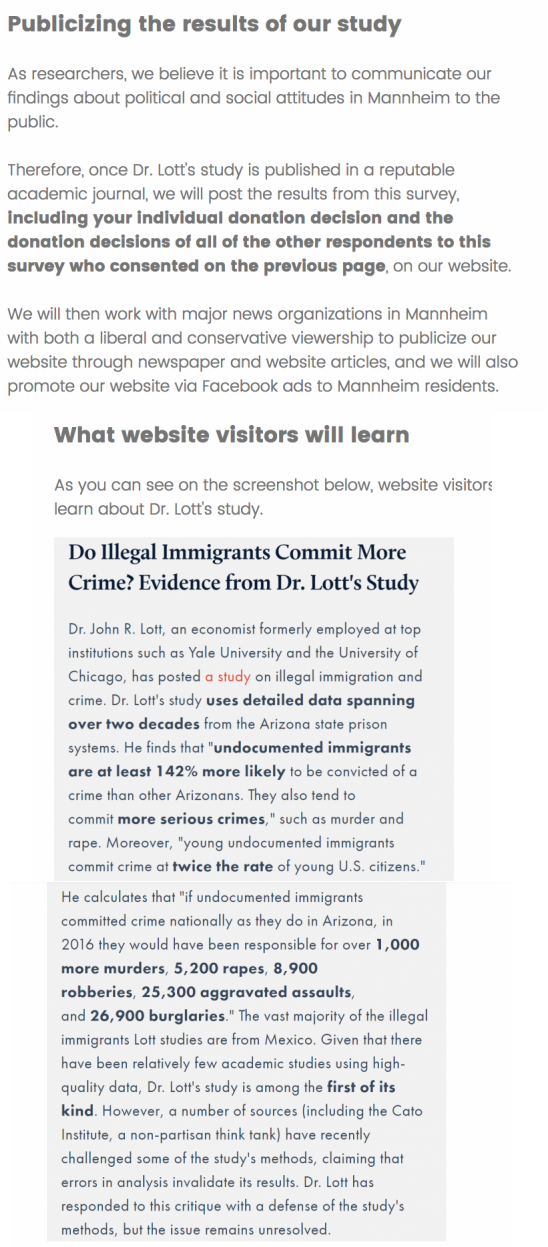


Description of donation decision





Website excuse condition and no excuse condition



Publicizing the results of our study

As researchers, we believe it is important to communicate our findings about political and social attitudes in Mannheim to the public.

Therefore, once Dr. Lott's study is published in a reputable academic journal, we will post the results from this survey, **including your individual donation decision and the donation decisions of all of the other respondents to this survey who consented on the previous page**, on our website.

We will then work with major news organizations in Mannheim with both a liberal and conservative viewership to publicize our website through newspaper and website articles, and we will also promote our website via Facebook ads to Mannheim residents.

What website visitors will learn

As you can see on the screenshot below, website visitors learn about Dr. Lott's study.

Do Illegal Immigrants Commit More Crime? Evidence from Dr. Lott's Study

Dr. John R. Lott, an economist formerly employed at top institutions such as Yale University and the University of Chicago, has posted a study on illegal immigration and crime. Dr. Lott's study **uses detailed data spanning over two decades** from the Arizona state prison systems. He finds that "**undocumented immigrants are at least 142% more likely** to be convicted of a crime than other Arizonans. They also tend to commit **more serious crimes**," such as murder and rape. Moreover, "young undocumented immigrants commit crime at **twice the rate** of young U.S. citizens."

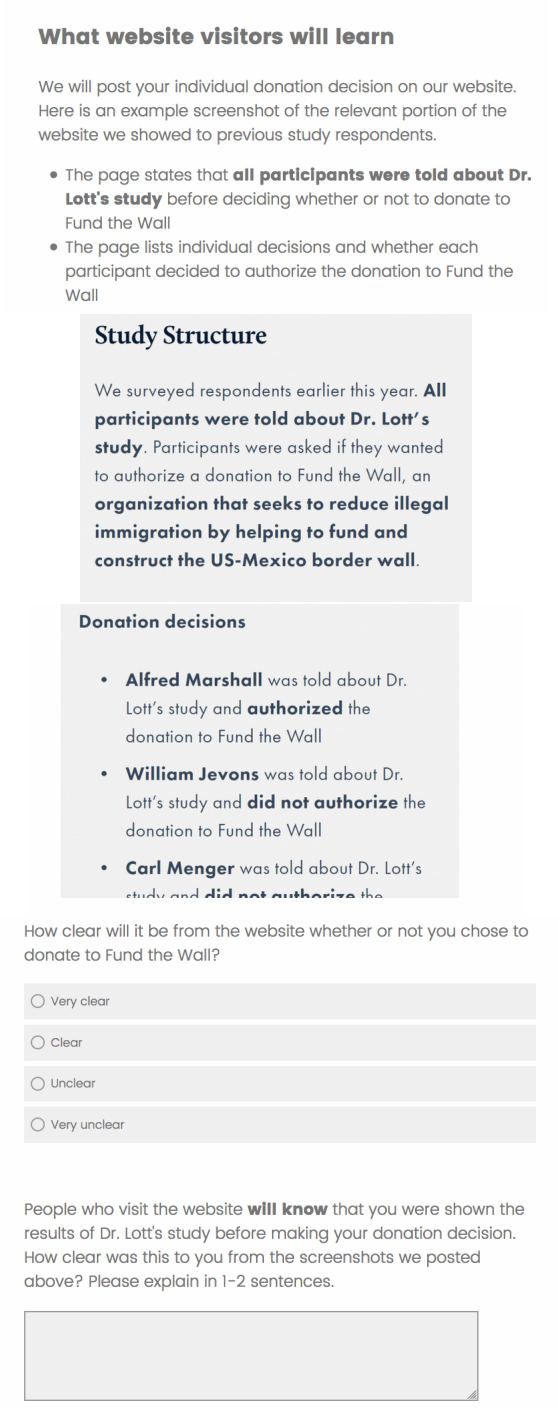
He calculates that "if undocumented immigrants committed crime nationally as they do in Arizona, in 2016 they would have been responsible for over **1,000 more murders**, **5,200 rapes**, **8,900 robberies**, **25,300 aggravated assaults**, and **26,900 burglaries**." The vast majority of the illegal immigrants Lott studies are from Mexico. Given that there have been relatively few academic studies using high-quality data, Dr. Lott's study is among the **first of its kind**. However, a number of sources (including the Cato Institute, a non-partisan think tank) have recently challenged some of the study's methods, claiming that errors in analysis invalidate its results. Dr. Lott has responded to this critique with a defense of the study's methods, but the issue remains unresolved.

What do you think website visitors will conclude about Dr. Lott's study after visiting the web pagepage?

- ○ That Dr. Lott finds that illegal immigrants are less likely to commit crimes than US citizens.
- ○ That Dr. Lott finds that illegal immigrants are equally likely to commit crimes as US citizens.
- ○ That Dr. Lott finds that illegal immigrants are more likely to commit crimes than US citizens.

>>

Donation: Excuse condition



Reminder: what website visitors will learn

- They will see that Dr. Lott's study finds that illegal immigrants commit more crime than US citizens
- They will learn whether or not you donated to Fund the Wall
- They will learn that **you were informed** about the results of Dr. Lott's study



Donation decision

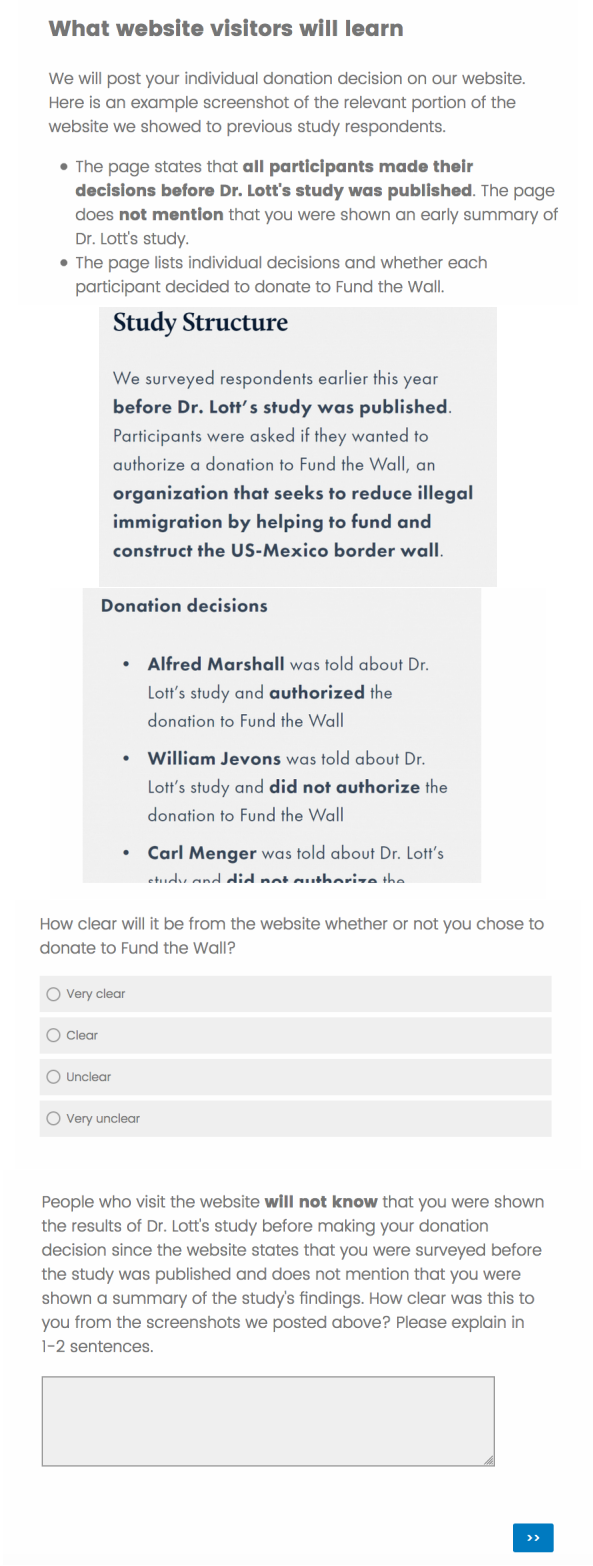
Would you like to authorize a $1 donation to **Fund the Wall**?

- ◯ Yes, I would like to authorize a $1 donation
- ◯ No, I would not like to authorize a $1 donation

Recall what people will learn when visiting the website:

- They will see that Dr. Lott's study finds that illegal immigrants commit more crime than US citizens
- They will learn whether or not you donated to Fund the Wall
- They will **learn that you were informed** about the results of Dr. Lott's study

Donation: No Excuse condition



Reminder: what website visitors will learn

- They will see that Dr. Lott's study finds that illegal immigrants commit more crime than US citizens
- They will learn whether or not you donated to Fund the Wall
- They will believe **you were not informed** about the results of Dr. Lott's study



Donation decision

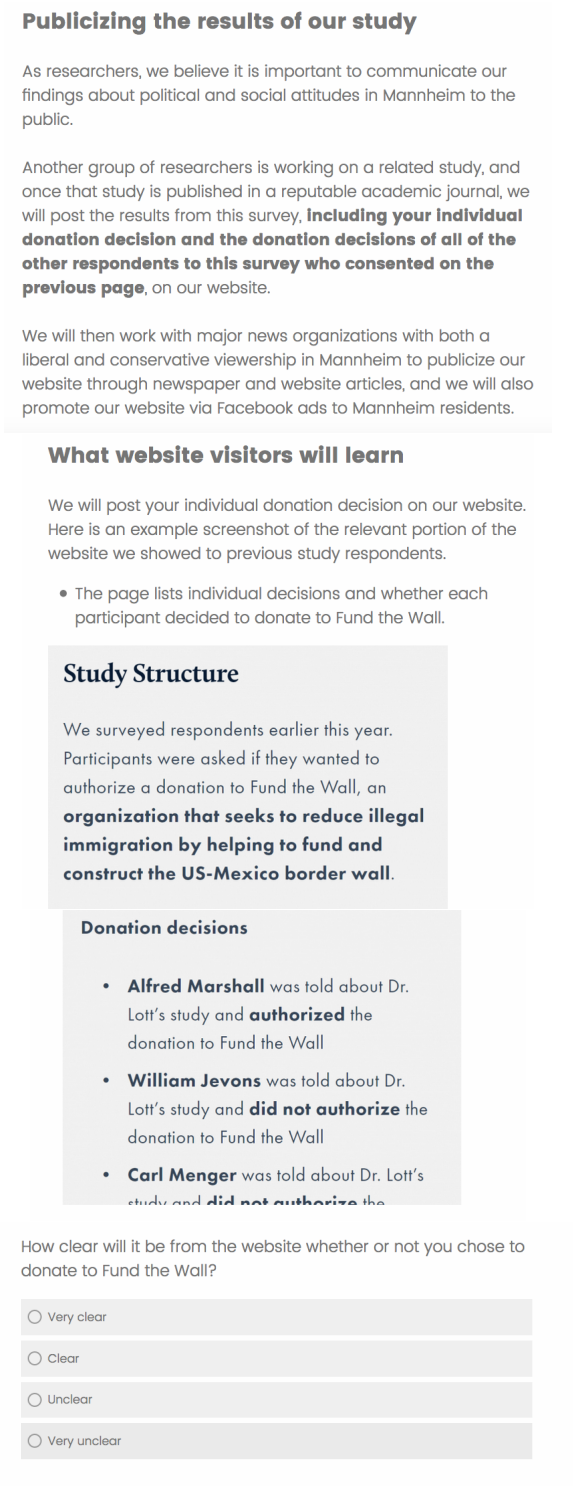
Would you like to authorize a $1 donation to **Fund the Wall**?

- ◯ Yes, I would like to authorize a $1 donation
- ◯ No, I would not like to authorize a $1 donation

Recall what people will learn when visiting the website:

- They will see that Dr. Lott's study finds that illegal immigrants commit more crime than US citizens
- They will learn whether or not you donated to Fund the Wall
- They will **not learn that you were informed** about the results of Dr. Lott's study, as the website states that you were surveyed before the study was published and does not mention that you were shown a summary of the study's findings

Donation: control condition



Publicizing the results of our study

As researchers, we believe it is important to communicate our findings about political and social attitudes in Mannheim to the public.

Another group of researchers is working on a related study, and once that study is published in a reputable academic journal, we will post the results from this survey, **including your individual donation decision and the donation decisions of all of the other respondents to this survey who consented on the previous page**, on our website.

We will then work with major news organizations with both a liberal and conservative viewership in Mannheim to publicize our website through newspaper and website articles, and we will also promote our website via Facebook ads to Mannheim residents.

What website visitors will learn

We will post your individual donation decision on our website. Here is an example screenshot of the relevant portion of the website we showed to previous study respondents.

- The page lists individual decisions and whether each participant decided to donate to Fund the Wall.

Study Structure

We surveyed respondents earlier this year. Participants were asked if they wanted to authorize a donation to Fund the Wall, an **organization that seeks to reduce illegal immigration by helping to fund and construct the US-Mexico border wall**.

Donation decisions

- **Alfred Marshall** was told about Dr. Lott's study and **authorized** the donation to Fund the Wall
- **William Jevons** was told about Dr. Lott's study and **did not authorize** the donation to Fund the Wall
- **Carl Menger** was told about Dr. Lott's study and **did not authorize** the

How clear will it be from the website whether or not you chose to donate to Fund the Wall?

- ○ Very clear
- ○ Clear
- ○ Unclear
- ○ Very unclear

Reminder: what website visitors will learn

- They will learn whether or not you donated to Fund the Wall



Donation decision

Would you like to authorize a $1 donation to **Fund the Wall**?

- ◯ Yes, I would like to authorize a $1 donation
- ◯ No, I would not like to authorize a $1 donation

Recall what people will learn when visiting the website:

- They will learn whether or not you donated to Fund the Wall

Post-outcome measures

