Disinformation Changes Minds and Closes Purse Strings for Charities

Erin Krupka[*] and Yi-Shan Lee[†]

July 26, 2022

Abstract

Americans like to donate to charities, but disinformation regarding either the charities or the recipients is a rising threat to this civic act. We examine the impact of disinformation on donation behavior, beliefs and social norms. Using two experiments on Amazon's Mechanical Turk, we find that exposure to a false negative claim regarding a subject (2018 migrant caravan or an endangered species) lowers the average donation towards it by 32%. We show that the false claim reinforces incorrect prior beliefs and changes what individuals consider the donation social norm to be. Although exposure to a false-claim seems to reinforce an incorrect prior, successfully correcting the belief by providing debunking information does not restore either behavior or norms to baseline levels. Our findings indicate that disinformation turns out to be a sticky threat – once present, it is difficult to undo its harm on behavior or social norms.

Key words: Charitable giving, Social norm, Information, Debunking, Fake news
JEL: C9, D64, D91

[*] School of Information, University of Michigan, 105 South State Street, Ann Arbor, USA.
[†]Corresponding Author. yishanlee@cuhk.edu.hk; CUHK Business School, Chinese University of Hong Kong, Shatin, Hong Kong. Yi-Shan Lee would like to acknowledge the hospitality of University of Michigan, where some of this work was carried out, and the financial support by the Swiss National Science Foundation (Doc. Mobility grant P1ZHP1_174630) and the Forschungskredit of the University of Zurich (Grant No. FK-17-017).

1. Introduction

Americans like to donate to charities, but disinformation regarding either the charities or the recipients is a rising threat to this civic act.[1] For example, both Save the Children and MSF (Medecins Sans Frontieres) were the subject of fake news in 2018 (International Broadcasting Trust, 2018), falsely accused of colluding with people traffickers as they conduct rescue efforts in the Mediterranean. In 2020, fake charities made the annual IRS Dirty Dozen tax scams report (Internal Revenue Service, 2020). The dissemination of "disinformation," the intentional spreading of inaccurate information with the desire to alter beliefs and attitudes, depresses civic life in many areas.[2] Previous literature focuses on how disinformation changes civic expression and participation, but less work has addressed the impact of disinformation on charitable contributions (DellaVigna and Kaplan, 2007; Martin and Yurukoglu, 2017; Pennycook et al., 2018; Lazer et al., 2018).

In this paper, we test the effect of disinformation on beliefs about those who would benefit from the charity and on charitable contributions. To unpack the channel through which false claims alter behavior, we build on a theoretical framework that includes a desire to comply with norms. Though others have pointed to motives for giving such as "warm glow", moral or social identity, social pressure, and social approval (Andreoni 1990; Ariely et al. 2009; Dellavigna et al., 2012), these motivations may be thought of as expressions of underlying norms. Hence, we focus on the role of social norms to donate. We take a utility-based approach (see also Allenby et al., 2020) which situates donation behavior in the context of utility maximization, wherein a donor considers marginal costs and benefits. In our case, the costs and benefits considerations include a desire to comply with social norms and disinformation alters those norms.

Our empirical approach is experimental. In our first experiment, we provide subjects with information regarding a real event related to the 2018 migrant caravan and then ask subjects to

[1] Nearly three-fourths of the US population participated in the civic act of donating to charity in 2019 – more than did in the presidential election of 2020.

[2] There are various types of inaccurate information, our study, depending the classification, is closely aligned with propaganda (persuasive information that aims to associate brands, people, products, or nations with certain feelings, ideas, and attitudes by mixing facts and interpretations with deliberate intent to manipulate or deceive (Jack, 2017; Tandoc, Lim, and Ling, 2017) or rumors (impressions, interpretations, or reactions that circulate but may not be true (Ruths, 2019).

decide how much they will donate to a charity that supports migrants. The experiment manipulates information exposure across four conditions: Baseline, False-claim, Debunk-before and Debunk-after. In the Baseline condition, subjects see a photo of the 2018 migrant caravan accompanied by the following neutral sentence: "A caravan with more than 1000 Central American migrants is currently traveling through Mexico, they are leaving their country after a contentious presidential election." In the False-claim condition, subjects see the photo of the caravan accompanied by the neutral statement and then see a second photo with the following false claim: "These throngs of Central Americans are trying to flood across our borders from Mexico so that they can take advantage of the United States DACA program." Debunking conditions present correcting (aka debunking) information before or after exposure to the false claim and allow us to rule out alternative channels by which disinformation impacts donation behavior. In addition, all conditions elicit incentivized beliefs about the DACA program, and use incentivized measures to elicit prescriptive and descriptive norms to donate. We also use self-reported measures to capture attitudes toward migrants.

A second experiment is conducted to test the robustness across contexts. Instead of the 2018 migrant caravan, we selected a non-political subject: the vampire ground finch, a small bird native to the Galápagos Islands, Ecuador. Using a parallel design of the four conditions, we construct a false claim that the birds were portrayed as migrating north, trying to cross U.S. borders so as to "take advantage" of US birds by pecking them to drink their blood. This statement is false as they are non-migratory birds. The structure of the experiment and incentivized outcome variables are exactly the same as in the first experiment.

Prescriptive social norms are defined as a collective perception among members of a population regarding appropriate behavior for a particular situation (Krupka and Weber, 2013). These norms are what a community believes one `ought' or `should' do. These are distinguished from descriptive social norms, which characterize what an individual believes most others are doing (Deutsch and Gerard, 1955; Bicchieri and Merceir, 2014; Bicchieri and Dimant, 2019). A long tradition in psychology distinguishes between prescriptive norms that convey information about

what one ought to do and descriptive norms which describe what is regularly done.[3] We follow this literature by focusing on prescriptive and descriptive norms.

The results of our study show that exposure to a false claim lowers overall donations by 32% in both experiments. The results from a hurdle-model analysis show that the impact on donations is driven mainly by a decrease in the number of individuals willing to donate (extensive margin) rather than by individuals choosing to lower their donation amount (intensive margin). Analysis of beliefs about the norm to donate, shows that the false claim weakens the prescriptive norm that prohibits donating nothing in the first experiment while weakens the descriptive norm in the second experiment. Thus, the norms and hurdle-model analysis yield consistent results – false claims weaken norms such that donating zero is more acceptable and increase zero donations, or lowering the expected donations from others and decrease one's own donation. Providing debunking evidence significantly improves individuals' belief accuracy regarding the false claim, however, correcting the false belief does not restore the impact of false claims on donation amounts or perceptions of social norms in both experiments.

This paper contributes to our understanding of the motives for acts of generosity by providing evidence that social norms to donate play an important role. Second, we document the profoundly negative role that a single piece of disinformation plays in eroding the norms to donate and their subsequent impact on behavior. Our debunking treatments show that disinformation regarding the recipients turns out to be a sticky threat—once present, it is difficult to undo its harm. Debunking false claims is a weak instrument by which to combat the degradation in donation norms introduced by the false claim. Finally, our study complements current research on the effects of retracting misinformation (Lewandowsky et al., 2012; Ecker et al., 2017) by expanding our understanding of the effect of debunking attempts to an examination of behavioral decisions and social norms.

Our study consists to two experiments designed to parallel each other on significant features: The migrant caravan study and the vampire bird study. In what follows, we will begin

[3] See also Cialdini et al. 1990; Krupka and Weber, 2009.

with a description of the migrant caravan study and results and then present the vampire bird study and results.

2. Experimental Design for the Migrant Caravan Study

We define four conditions that exogenously vary whether a subject is or is not exposed to a false claim and whether they are exposed to debunking information before or after the false claim. These treatments are outlined in Table 1.

This study was conducted using Amazon's Mechanical Turk in May 2018. For our news event, we used the march of Central American migrants through Mexico to the United States border, the so-called "2018 caravan." Regardless of treatment, the experiment has four parts.

Table 1: Experimental Conditions Migrant Caravan Study

Condition	False Claim	Debunking	Timing of Debunking
Baseline	No	No	-
False-claim (FC)	Yes	No	-
Debunk-before (D+FC)	Yes	Yes	Before seeing false claim
Debunk-after (FC+D)	Yes	Yes	After seeing false claim

In part one, we elicit participants' prior beliefs about whether migrants are eligible for DACA through two multiple-choice questions. The first question asks participants to guess "the number of people who entered the US in 2016 that were eligible for DACA"[4] by choosing one answer from the following options: zero, 10^3, 10^4, 10^5 and 10^6. This question is incentivized with a bonus of a \$0.25USD for the correct response of "zero."[5] To verify that a correct response to the first question is not a random guess, we include a second question that provides a neutral description of the marching group: "A caravan with more than 1000 Central American migrants is currently traveling through Mexico, they are leaving their country after a contentious presidential election." We ask participants to check each statement that represents the actions they believe

[4] Asking the eligibility for migrants in 2016 enables us to elicit subjects' understanding about DACA eligibility while avoiding linking these prior belief questions directly to the 2018 caravan in the information exposure.
[5] Figure A.1 in Appendix summarizes the proportions and numbers of participants with various prior beliefs in each condition.

migrants in the caravan could take. The options are: "Some will reach Mexico and legally seek asylum and legally stay in Mexico"; "Some will reach the US border and legally seek asylum and legally stay in the US"; "Some will cross the US border and use DACA to legally stay in the US"; and "Some will cross the US border and stay illegally in the US." Subjects can also write in other options through the option "Others." A correct response should not include the option "Some will cross the US border and use DACA to legally stay in the US."[6]

In part two, subjects see different content depending on their experimental treatment. In our baseline, subjects view the photo of the caravan with the accompanying statement: "A caravan with more than 1000 Central American migrants is currently traveling through Mexico, they are leaving their country after a contentious presidential election." After exposure to this statement, they are asked to paraphrase the statement as an attention check. In the false-claim condition, all aspects mirror the baseline, except that after paraphrasing the statement about the caravan, subjects see an additional picture of the caravan with an extra accompanying statement. They are asked to paraphrase the accompanying statement as well. The statement reads: "Claim: These throngs of Central Americans are trying to flood across our borders from Mexico so that they can take advantage of the United States DACA program." [7] The claim is factually false because DACA requires immigrants to be "physically present in the United States on June 15, 2012," making the migrants in the caravan ineligible for the program. However, it also contains an unverifiable negative claim which characterizes the motivations ("so that they can") of migrants.

The design of the two debunk conditions (debunk-before and debunk-after) is identical to the false claim treatment, except that we show subjects a screenshot of DACA eligibility guidelines from the US Department of Homeland Security website before or after (depending on the treatment) being exposed to the false claim. On the screenshot, the eligibility criterion of being "physically present in the United States on June 15, 2012" is highlighted. Accompanying the screenshot is a

[6] This question is not incentivized. Participants are only paid if they get the first question correct.
[7] This study addresses the potential perception of deception by identifying it as a claim and providing the correction at the end (in the case of the false claim treatment) or just prior or after the claim (in the debunking treatments).

short paragraph that explains the migrants in the caravan photo do not satisfy the highlighted condition and explicitly informs participants that the false claim is indeed false.[8]

The third part of our experiment is identical for all treatments. We elicit four incentivized outcome measures. Subjects have the opportunity to donate up to $1 to a fundraiser organized by Undocumedia, a 501c3 nonprofit organization in Los Angeles, CA and Pueblo Sin Fronteras, an international migrant and refugee rights collective, to support the 2018 Refugee Caravan.[9,10] Subjects are asked to select a donation amount between zero and one dollar, in 10 cent increments. Comparing differences in donation amounts across conditions allows us to observe the causal effect of false claims and debunking information on the donation.

The second incentivized measure asks participants to estimate the percentages of other participants who donate each listed amount, to measure their beliefs about the donation distribution. Specifically, we show them a list that ranges from $0.00 USD to $1.00 USD in 10 cent increments. Next to each 10 cent increment, they indicate the percentage (out of 100) who they believe donated that amount. All percentages should add up to 100 and the program does not allow them to advance until they do. We provide incentives to guess this distribution correctly. One guess (e.g., their guess regarding the proportion who donated 20 cents) is randomly selected. If the guessed proportion is within 5 percentage points of the actual proportion who donated that amount, then they receive 50 cents. This incentivized first-order belief measure captures beliefs about the descriptive norm of donating among our participants.

[8] See Appendix 1 for experimental instructions.

[9] We include a small paragraph to describe the caravan: "These migrants consist of unaccompanied minors, women migrating on their own, and family units. Your donation will be "directly administered by migrant leaders and volunteers from Pueblo Sin Fronteras who are accompanying the caravan through Mexico". At the conclusion of our study, we add all the donations together and send them to Undocumedia and submit the donations as an anonymous donor. At the end of this study, we also provide them with a link to the fundraising website for Immigrants in 2018 Refugee Caravan so they can learn more about the organization."

[10] According to the website of Pueblo Sin Fronteras (English: Village without Borders), Pueblo Sin Fronteras is "a transborder organization made up of human rights defenders of diverse nationalities and immigration statuses that promotes accompaniment, humanitarian assistance, leadership development, recognition of human rights, and coordination of know-your-rights training along migrant routes, as well as monitoring and raising awareness of human rights abuses against migrants and refugees in Mexico and the United States."

In our third incentivized measure, we ask participants to indicate the appropriateness of a specified donation amount, to obtain an empirical proxy for the social norm of donating (Krupka and Weber, 2013). For each donation amount, we ask whether making that donation amount would be "very socially inappropriate," "somewhat socially inappropriate," "somewhat socially appropriate," or "very socially appropriate." Following Krupka and Weber (2013), we further explain that by socially appropriate, we mean that most Mturkers would consider the action to be consistent with moral or proper social behavior". To ensure incentive compatibility, we randomly select an appropriateness rating; participants have the possibility of earning $0.50 if their choice is the same as the most common appropriateness rating provided by other participants in that treatment.

In our fourth incentivized measure, we ask participants to indicate the number of migrants in the 2018 caravan eligible for DACA. In the debunking conditions, this allows us to test if debunking leads to a more accurate belief about how the policy applies to the 2018 caravan. A participant receives an additional $0.25 payment for the correct answer (which is "zero").

Finally, we conclude our study with a questionnaire to collect participant demographic and ideology data. We collect gender, educational level, age, residential state in the US, political ideology ("Democratic", "Republican"," Independent" or "Other"), ideology on social issues ("Strongly Liberal", "Somewhat Liberal", "Moderate", "Somewhat Conservative", or "Strongly Conservative"), personal feelings about the nature of the event ("positive", "negative" or "no strong feelings"), and pre-experiment knowledge about the event. Participants in the treatment conditions (false-claim, debunk-before, and debunk-after) are also asked about the extent to which exposure to the claim affected their donation decision as well as any pre-experiment exposure to the information about the 2018 caravan and the claim.

To obtain our sample, we recruit subjects through Amazon's Mechanical Turk. We require subjects to be US citizens at least 18 years of age. Our total sample consists of 711 participants who are assigned to one of six sessions, with 184 participants in the Baseline condition, 177 in False-claim, 172 in Debunk-before, and 178 in Debunk-after.

Table 2: Balance of Covariates in the Migrant Caravan Study

	Baseline	False-claim	Debunk-before	Debunk-after	p-value
	(1)	(2)	(3)	(4)	(5)
Female	0.402	0.435	0.494	0.399	0.243
	(0.492)	(0.497)	(0.501)	(0.491)	
Education	3.755	3.785	3.785	3.815	0.880
	(0.669)	(0.698)	(0.680)	(0.701)	
Age	34.870	35.028	36.145	35.382	0.687
	(9.588)	(10.80)	(10.51)	(11.87)	
Republican	0.196	0.192	0.174	0.174	0.927
	(0.398)	(0.395)	(0.381)	(0.380)	
Heard News	0.473	0.441	0.430	0.427	0.810
	(0.501)	(0.498)	(0.497)	(0.496)	
N	184	177	172	178	

Notes: Columns (1) to (4) report the mean level of each variable for each of our four conditions, with standard deviations in parentheses. Column (5) indicates the p-value when testing that means are the same across conditions. The variables "Female," "Republican," and "Heard News" are dummy variables. "Heard News" equals one if for subjects with pre-experiment exposure to the information about the 2018 caravan. Education is an ordinal variable coded with following specifications: 1: Elementary school, 2: Junior high school, 3: Senior high school, 4: Undergraduate school (College), and 5: Graduate school (Masters or professional).

We conduct the experiment on Qualtrics, which is also used to randomly assign subjects in each session to four conditions. The average participation time is about 12 minutes. Table 2 outlines our conditions by subject and shows balanced individual characteristics across all four conditions. To be specific, columns (1) to (4) report the mean levels of each characteristic of our participants for four conditions. Column (5) indicates the p-values when testing that means are the same across conditions. Based on the statistics in Table 2, we conclude that our randomization procedure is successful.[11]

2.2 Hypotheses

To test the effect of our treatments on behavior and beliefs, we form the following hypotheses about the expected change in our outcome measures across conditions. We predict that participants who are exposed to the false claim regarding DACA eligibility will donate less.

[11] Here we report only personal characteristics. Figure A.1 reports the distribution of prior beliefs about the DACA policy for our four conditions. Rank-sum tests confirm that the distributions of prior beliefs about the DACA policy in all three experimental conditions are not significantly different from those in the Baseline group.

Hypothesis 1 (Donations): *The average donation will be lower in the false claim condition than in the baseline.*

The first potential mechanism is that the impact of a false claim may operate on beliefs by reinforcing an incorrect prior belief.[12] Reinforcement has been shown to lead to an increase in the likelihood of voluntarily exhibiting that behavior in response rather than an increase in the intensity with which a behavior is engaged in (Blough and Millward, 1965; Skinner, 1937). Thus, a reinforcing pathway, suggests that exposure will lower donation probabilities for those with incorrect priors. Second, reinforcement learning predicts that the false claim will affect the probability of donating rather than the amount donated.

Hypothesis 2a (Donation probability: Beliefs reinforced): *The probability of donating will be lower for those with incorrect priors.*

Hypothesis 2b (Conditional donation amount: Beliefs reinforced): *Conditional on making a donation, the average donation will not differ between the baseline and false claim condition.*

Yet a different pathway may be that norms of supporting migrants are impacted. We might assume that actor i takes action a_i. And further, that the actor has a belief function, $\delta_i=E(a_{-i})$, which assigns to each possible action (available to a decision maker in that situation) a belief about the frequency the action is taken by others. We call this subjective belief the descriptive norm. We also imagine that each actor has a prescriptive norm, η_i, which is a function that assigns to each action a degree of appropriateness or inappropriateness and, more conventionally, is articulated as what an agent *ought* to do.

We assume the agent's utility from taking action a_i has three components - the direct utility from the action ($U^* - \alpha_i(a_i - a^*)^2$), the disutility from violating one's prescriptive norm ($\beta_i(a_i - \eta_i)^2$) and the disutility from taking an action that differs from one's descriptive norm ($\gamma_i(a_i - \delta_i)^2$):

[12] An alternative is that false claims change behaviors by changing beliefs about DACA. Perhaps surprisingly, we only have three observations that change from a correct prior to an incorrect posterior belief after reading the false claim. Thus, this pathway, for this setting, does not seem likely.

$$U(a_i) = U^* - \alpha_i(a_i - a^*)^2 - \beta_i(a_i - \eta_i)^2 - \gamma_i(a_i - \delta_i)^2$$

Here a^* denotes the agent's ideal donation without concerns for norms. The quadratic utility loss ensures that the agent's direct utility decreases the further away her action is from the ideal action. Similarly, the norm violation disutilities increase with the distance of the action from the respective norm. This model is a quadratic version of the model in Krupka and Weber (2013) who allow for a more general norm utility/disutility function $N(.)$. One advantage of the quadratic formulation is that it is easy to derive the agent's optimal action:

$$a_i = \frac{\alpha_i a^* + \beta_i \eta_i + \gamma_i \delta_i}{\alpha_i + \beta_i + \gamma_i}$$

Intuitively, the agent's optimal action is a weighted average of her ideal action, the prescriptive norm and the descriptive norm. The sensitivity to norm compliance is captured by the weights $\beta_i \geq 0$ and $\gamma_i \geq 0$. The larger the sum of these weights, the more the agent cares about norm compliance. In our setting, these norms are the appropriateness/empirical expectation of donating.

False claims that couple factual (though debunkable) statements with aspersions to intentions (not debunkable) may make migrants less deserving of help and either change the prescriptive or descriptive norm (η or δ) or weaken concern (β or γ) to comply with a donation norm (Deutsch and Gerard, 1955). We can test these channels. If the false claim changes the descriptive or prescriptive norm, then we can construct the following hypotheses:

> ***Hypothesis 3a (Mechanism: Descriptive norm change):*** *Subjects believe that the average donation (descriptive norm) made by others is lower in the False-claim condition than in the baseline.*

> ***Hypothesis 3b (Mechanism: Prescriptive norm change):*** *Subjects believe that it is more socially appropriate (prescriptive norm) to donate zero in the False-claim condition than in the baseline.*

In addition, if the norm for donating zero has changed, such that it is more appropriate in the false claim condition, then we predict more zero donations.

> ***Hypothesis 3c (Mechanism: Norm change):*** *There will be more zero donation choices in the false claims condition than in the baseline.*

Continuing with this line of reasoning, we can use a conditional logit to test whether norms predict donating and whether false claims reduce the weight a participant attaches to an prescriptive or descriptive norm of donating to the migrants.[13] Note that in a conditional logistic regression (McFadden, 1973) where the dependent variable is the choice alternatives, the independent variables are attributes of the choice alternatives. The coefficient on each independent variable can be thought of as the average weight that subjects attached to each attribute of the choice (payoff, prescriptive and descriptive norm characteristics in our situation). A significant interaction of norms and our treatment variable can be interpreted as a change in the average weight given to consideration of the norm in the false-claim condition. This leads to the following hypotheses:

> ***Hypothesis 3d (Mechanism: Concern for norm compliance predicts behavior):*** *The coefficient on norms, will be a positive and significant predictor of donations; a model that includes norms will improve our ability to explain behavior (as measured through decreases in AIC/BIC) than one that does not.*
>
> ***Hypothesis 3e (Mechanism: False claims weaken concern for compliance):*** *The interaction on norms and treatment will be negative and significant denoting a reduction in the weight an individual attaches to norm compliance in the false claim treatment.*

Depending on the channel through which debunking works, the impact of debunking can provide further empirical support regarding the channel through which false-claims operate. Debunking may mitigate the impact of a false claim on donations by correcting the reinforced false belief or by buffering the negative impacts of the false claim on norms. This yields the following hypotheses:

> ***Hypothesis 4a (Successful debunking):*** *Exposure to debunking information leads to more accurate beliefs about DACA eligibility compared to those who receive a false claim without debunking information.*

[13] Numerous papers provide evidence that there is heterogeneity in individual concern for complying with a given norm. These papers demonstrate that people can articulate what the norm is but that some will care a great deal, while others very little, about compliance with that norm. In Krupka and Weber (2013) they articulate a utility model in which the utility of taking a certain action depends on the social appropriateness and monetary payoffs, and then use a conditional logit to estimate the average weight placed on norm compliance. In Kimbrough and Vostroknutov (2016), they devise a method for estimating the individual weights placed on norm compliance (and take the norm as given). They show that this heterogeneity is predictive of norm compliance.

Hypothesis 4b (Debunking: Recovery in donation): *The average donations of those in debunking conditions will not significantly differ from the baseline donation; average donations in debunking conditions will be greater than in the false claim condition.*

Hypothesis 4c (Debunking: Beliefs reinforced): *Conditional on having incorrect priors, the average donations of those who update to the correct posterior belief after debunking will be higher than those who retain their incorrect priors.*

Hypothesis 4d (Debunking: Recovery in norms): *The norm in debunking conditions will not significantly differ from the baseline donation.*

We now turn to testing these hypotheses.

3. Results

The shaded bars in Figure 1 show the average donations across the four conditions. From the first two bars, we see that the average donation in the false-claim condition is significantly lower than that in the baseline condition.

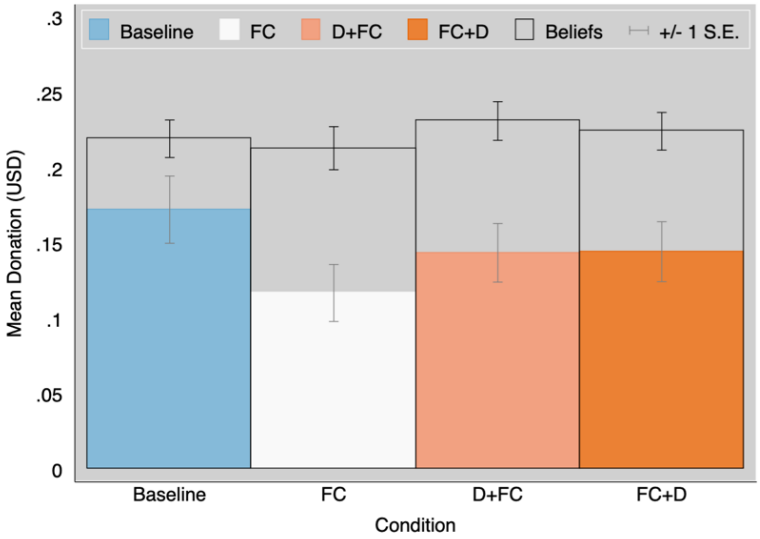

Figure 1: Mean Levels of Donations (shaded bar) and Mean Beliefs Regarding Other's Donations (framed bar) out of 1 USD.

Individuals exposed to a false claim significantly decreases the average donation toward migrants by 32%, from \$0.172 of the available dollar to \$0.117 (p-value<0.05, one-tailed t-test).

The results from a Wilcoxon-Mann-Whitney test reinforce this finding as they allow us to reject the conjecture that the False-claim donation distribution is equal to that of the Baseline (p-value<0.05). From these results, we conclude that exposure to a false claim regarding the migrant caravan negatively impacts participants' decisions to make a donation to an organization supporting the migrants.[14]

> ***Result 1 (Donations):*** *The average donation in the false-claim condition is 32% lower than the average donation in the baseline condition (p<0.05).*

We now test whether the mechanism by which false claims operate is to reinforce an incorrect prior belief.[15] We test this conjecture in two steps.

Table 3: Hurdle Model Estimation on Donation

	Full sample	**Only those with incorrect priors**
False-claim: hurdle	-0.257^{+}	-0.298*
	(-1.87)	(-2.03)
False-claim	-0.118	-0.104
	(-0.74)	(-0.56)
Average marginal effect	-0.055*	-0.056*
	(-1.99)	(-2.06)
N	361	319

Notes: The z-statistics are reported in parentheses. The average marginal effect of the false claim on donation is reported for each model as the coefficient estimates are not directly interpretable. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Using a hurdle model proposed by Cragg (1971), we test whether false claims significantly decreased donation probability (the extensive margin) or reduce the donation amount conditional on

[14] Most participants were unaware of this effect. As shown in Appendix 3, 116 out of 177 participants in the false-claim condition self-reported that seeing the false claim has "none at all" impact on their donations in our questionnaire.
[15] A correct prior belief means that a subject answers "zero" to the question "guess the number of people who entered the US in 2016 that were eligible for DACA." Choosing a non-zero answer indicates that the subject has an incorrect prior and may not be able to notice the infeasibility in the false claim.

donating (the intensive margin).[16] Results are summarized in Table 3. This model assumes that individuals first decide whether they will donate; those who choose to donate then decide on their donation amount. The two-step decision is modeled as a probit regression predicting the donation decision and a linear regression predicting the donation amount.

Column one shows the results when we include all subjects (those with correct and incorrect priors). We find directional evidence that exposure to the false claim reduces the donation probability relative to the baseline but this result is marginally significant ($z=-1.87$, p-value<0.10). Conditional on donating, there is no difference in the amount donated ($z=-0.74$, n.s., column 1).[17] Column two runs the same regression, but only for those subjects who have incorrect priors. We see that the donation probability is lower ($z=-2.03$, p-value<0.05) and, conditional on donating, there is no difference in the amount donated ($z=-0.56$, n.s., column 2). From these results, we conclude that the effect of false claims on behavior happens mainly at the extensive margin in our context. This result supports the hypothesis 2a and 2b that exposure to the false claim reinforces an incorrect prior belief such that subjects in the false-claim condition are less likely to make a donation.

> ***Result 2 (Mechanism: Beliefs reinforced):*** *The hurdle model finds evidence consistent with reinforcement models: A false claim lowers the probability of donating in the false claim treatment (relative to the baseline). This effect is pronounced among those with incorrect priors ($p<0.05$). However, conditional on donating, there is no difference in the amount donated. Taken together, these results support a mechanism whereby false claims are reinforcing an incorrect prior belief.*

We now turn to testing hypotheses 3a-3e regarding the impact of exposure to a false claim on what people consider appropriate to do— i.e. testing if the norm is changed. Table 4 presents regressions testing the first three hypotheses.

[16] Figure A.2(a) provides a histogram of the donation adjustment results for the Baseline and False-claim conditions. As seen from the distribution, exposure to the false claim False-claim largely increases the proportion of participants who choose not to donate.

[17] Although our coefficient estimates are not directly interpretable, a post estimation shows that exposure to the false claim lowers the donation by about \$0.05 ($z=-1.99$, p-value<0.05), which is 29% of the baseline donation level. Wilcoxon-Mann-Whitney tests also show similar results regarding the comparison between donation distributions in the false-claim and baseline conditions. Specifically, these results allow us to reject the conjecture that the baseline and false-claim donations are from the same distribution ($z=1.99$). However, the donation distribution is not significantly different when it is conditional on one being a donor ($z=0.79$).

Column one finds no significant differences between treatments regarding beliefs about what others will donate (beta=-0.005, n.s.).[18] However, column two shows that exposure to a false claim significantly increases the socially appropriateness of donating zero (beta=0.263, p<0.05). On average, individuals consider the choice of donating zero "very socially inappropriate" in the baseline condition, but only "somewhat socially inappropriate" in the false-claim condition.[19] As the prescriptive norm of donating zero has changed, hypothesis 3c tests whether this is correlated with an increase in zero donations. Column three finds support for this: There are more zero donations after exposure to the false-claim. The result is significant at 5% level after controlling for prior belief, education, age, gender and political ideology.

Table 4: The Effect of Exposure to the False Claim on Social Norms and Donation

DV	(1) OLS Descriptive norm ratings	(2) OLS Prescriptive norm for donating zero	(3) LOGIT Donate zero=1 Marg. Effects
False claim	-0.005	0.263*	0.097*
	(-0.28)	(2.34)	(2.01)
+Controls	Yes	Yes	Yes
_cons	0.208***	1.939***	
	(3.44)	(5.41)	
N	361	361	361

Notes: The control variables include education, age, gender, political orientation and the prior belief of DACA eligibility. The *t*-statistics are reported in parentheses; marginal effects are reported for logit regressions. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The conditional logit estimates in Table 5 present evidence that the prescriptive norm predicts donations (beta=0.914, p-value<0.01, col. 2) while the interaction term shows that the weight a participant attaches to an prescriptive norm does not differ by treatment (beta=-0.347, n.s., col. 3). In addition, the comparison between the BIC (Bayesian information criterion, which

[18]A Wilcoxon-Mann-Whitney test further shows that the expected donation distribution in the False-claim condition does not significantly differ from that in the Baseline (p-value=0.363, n.s.). These test statistics rule out the possibility that the change of donation behavior is driven by the expectation that others may donate less in the presence of false-claim. Hence, exposure to the false claim does not change the descriptive norm of donating to migrants.

[19] Two-sample test of proportions shows the proportions of subjects consider not donating more appropriate (choosing very socially appropriate or somewhat socially appropriate) is significantly higher in the False-claim condition than in the Baseline (p-value<0.01, two-tailed).

penalizes model complexity) of models one and two shows that including the normative norm rating when predicting the chosen action of donation increases the model's predictive fit.[20]

> ***Result 3 (Mechanism: Prescriptive norm change):*** *In comparison with the baseline, false-claim exposure causes the prescriptive norm to weaken for donating zero, making it more acceptable, and causes an increase in zero donation amounts. Regression estimates suggest this behavioral change is not driven by changes in the weight placed on norm following between baseline and false claim.*

Table 5: Conditional Logit on the Prescriptive Norm and Chosen Donation Amount

	(1) chosenAction	(2) chosenAction	(3) chosenAction
Payoff to Self	5.560***	7.132***	7.247***
	(4.55)	(5.41)	(5.45)
Appropriate		0.914***	1.071***
		(5.25)	(5.49)
Appropriate x False Claim			-0.347
			(-1.71)
+Controls	Yes	Yes	Yes
AIC / BIC	922.079/1242.704	892.145/1219.057	891.174/1224.372
N	3971	3971	3971

Notes: The control variables include education, age, gender, political orientation and the prior belief of DACA eligibility. *t* statistics in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Results two and three test for two channels through which a false claim impacts behavior: It can reinforce incorrect priors and change what subjects consider to be socially appropriate (the prescriptive norm of donation). We now investigate our debunking treatment for further evidence in support of these channels.

We first show that subjects with inaccurate priors update beliefs in our debunking conditions. We find that 32.56% (45.51%) correct their beliefs about DACA eligibility in the before (after) debunking treatment and the proportion of subjects updating is significantly different from that of the false-claim and the baseline condition (two-sample test of proportions p-value<0.001 for

[20] Both penalize-likelihood criteria: AIC and BIC are reported. We focus on BIC as its penalty term for adding parameter is larger and is thus better in situations where a false positive is as misleading as, or more misleading than, a false negative.

all four pairs of comparisons, z=-7.59, -9.54, -7.74, -9.72).[21] Note that we do not find a difference in the proportion with correct prior beliefs in debunking conditions from conditions without debunking (z=-0.61).

> ***Result 4a (Mechanism: Beliefs changed):*** *Exposure to debunking information leads to more accurate beliefs about DACA eligibility compared to those who receive a false claim without debunking information.*

However, as we can see in Figure 1, debunking is only partially effective in resorting donation behavior to baseline levels by increasing donations. The average donation after debunking ($0.14 for both debunking conditions) hovers between levels observed in the baseline and false claim condition, neither significantly higher than the mean donation in the False-claim condition ($0.12) nor significantly lower than the baseline ($0.17).[22] Even when we only consider those who update in the debunk before/after treatment (33% and 46%, respectively), we find that these participants do not donate more than those who do not update their beliefs (p-value=0.480 in debunk-before and p-value=0.698 in the debunk-after).[23] Lastly, we find that subjects who accurately update their beliefs after debunking do not donate more than participants in our false-claim condition (p-value=0.351 in the debunk-before and p-value=0.515 in the debunk-after condition). These results suggest, but are not conclusive, that successful debunking cannot ameliorate the negative impact of false claims on donations and lend further support to the notion that false claims alter more than beliefs.

[21] In our experiment, we are also interested in whether the timing of exposure to debunking information impacts the effect of this information on the subsequent donation decision. As previous literature suggest that preexposure warning may limit the effect of misinformation compare to retract afterward (Chambers & Zaragoza, 2001; Ecker et al., 2010; Jou & Foreman, 2007; see Lewandowsky et al., 2012 for a great review), we vary whether subjects are exposed to debunking information before or after exposure to the false claim. Our results show the proportion of subjects who correctly update their beliefs in the debunk-before treatment is significantly different from (lower) that in the debunk-after at 5% level (z=-2.36).

[22] Notice that the power for this one-sided test with a 5% significance level and the given means and standard deviations is 60%. Since the respective sample sizes of the two Debunking conditions (172 and 178) do not differ much from those of the Baseline and False-claim conditions (184 and 177, respectively), it is more likely that the power is lower than the conventional 80% level due to large standard deviations.

[23] The results unconditional on holding a false prior belief are even more similar: p-value=0.745 (0.823) in Debunk-before (-after).

Result 4b (Donation: Beliefs changed)*: The average donation of those in debunking conditions is not significantly more than the false claim condition and not significantly different from the baseline.*

Result 4c (Debunking: Beliefs reinforced)*: Conditional on having incorrect priors, the average donations of those update to the correct post belief after debunking do not differ from those retain their incorrect priors.*

When we examine the impact of debunking on norms, we find evidence that supports the interpretation that false claims change norms. The average norms rating of subjects in the debunking conditions who updated their incorrect prior to a correct posterior is 2.18. That is, even in the debunking treatments, and even when they correct their false priors, subjects view donating zero as more appropriate than those in the baseline and this difference is significant (p-value<0.05, two-tailed).[24] This result also holds when we limit the baseline sample to individuals with incorrect priors (p-value<0.05, two-tailed). Thus, even successful debunking cannot reverse the damaging impact of a false claim on norms to donate.

Result 4d (Debunking: Fail to recover prescriptive norms)*: Individuals who experienced successful debunking in the debunking conditions still consider choosing not to donate more socially appropriate. Thus, hypothesis 4d is not supported.*

In aggregate, results 4a-4d support the interpretation that false-claim exposure causes changes in behavior by changing the prescriptive norms. This effect is irreversible even after successful debunking. Although exposure to a false-claim seems to reinforce incorrect prior, correcting the false belief does not alleviate the negative impact of false-claim on behavior. Overall, our results suggest that alleviate the impact of false-claim on behavior needs more than debunking.

Experiment 2

To test the robustness of our results and interpretations about the underlying mechanisms, we ran a second experiment with the same design as in the first experiment but using a different context. Instead of using the 2018 migrant caravan, which received wide news coverage and was

[24] The average appropriateness measure is 2.05 (2.11) in debunk-before (-after) and the pooled average appropriateness measure across the two debunking conditions is 2.08.

picked up by politicians, we invented an event that has no political or media coverage. We used a real bird, the vampire ground finch, which is a small bird native to the Galápagos Islands, Ecuador. We constructed a parallel scenario to that of the migrant caravan in that the birds were portrayed as migrating north, trying to cross U.S. borders so as to "take advantage of US birds" by pecking them to drink their blood. This is a false statement as they are non-migratory birds. The four conditions are the same as those outlined in Table 1. This follow-up study was conducted using Amazon's Mechanical Turk in September 2021.

In part one, all subjects answered two multiple choice questions which elicit prior beliefs about whether vampire birds can take advantage of the natural resources and birds found in the U.S.. Specifically, the first question asks participants to guess "the number of vampire birds that migrated northward, crossing the U.S. border, in 2016?".[25] They can choose one response from the following options: zero, 10^3, 10^4, 10^5 and 10^6. This question is incentivized with a bonus of a \$0.25USD for the correct response of "zero."[26] To verify that a correct response to the first question is not a random guess, we include a second question that provides a neutral description of the species[27] and asks participants to check each statement that represents the actions they believe the birds could take "if there is food scarcity where they live". The options are: "Some will drink the blood of nesting birds"; "Some will eat other invertebrate insects"; "Some will fly to the US border and come into the US to drink the blood of US nesting birds"; "Some will fly to other parts of the island and drink the blood of nesting birds." Subjects can also write in other options through the option "Other." A correct response should not include the option "Some will fly to the US border and come into the US to drink the blood of US nesting birds."[28]

In part two, subjects see different content depending on the experimental treatment they are randomized to. In the baseline, subjects view the photo of the vampire ground finch with the accompanying statement: "The vampire ground finch (Geospiza septentrionalis) is a small bird native to the Galápagos Islands, Ecuador. Their home is just south of Mexico and the US. When

[26] Figure A.1(b) in Appendix summarizes the proportions and numbers of participants with various prior beliefs in each condition.
[27] The description was: "The vampire ground finch (Geospiza septentrionalis) is a small bird native to the Galápagos Islands, Ecuador. Their home is just south of Mexico and the US. When food is scarce, the vampire finch occasionally feeds on the blood of other birds by pecking them."
[28] This question is not incentivized. Participants are only paid if they get the first question correct.

food is scarce, the vampire finch occasionally feeds on the blood of other birds by pecking them." After exposure to this statement, they are asked to paraphrase the statement as an attention check.

In the false-claim condition, all aspects mirror the baseline, except that after paraphrasing the statement about the species, subjects see an additional picture of the bird with an extra accompanying statement. They are asked to paraphrase the accompanying statement as well. The statement reads: "Claim: These throngs of vampire birds from Ecuador are trying to cross our borders into the United States to take advantage of US birds by pecking them to drink their blood." The claim is factually false because the vampire ground finch is not a migrant. Notice that in this experiment we no longer have an unverifiable negative claim about the motivation. This adjustment helps us rule out the possibility that the long-lasting effect of the false claim, if any, was driven by that un-debunkable claim.

The design of the two debunk conditions (debunk-before and debunk-after) is again identical to the false claim treatment, except that we show subjects a screenshot from the International Union for Conservation of Nature (IUCN) Red List of Threatened Species showing that the vampire ground finch is "not a migrant" before or after (depending on the treatment) being exposed to the false claim. On the screenshot, the bird is clearly identified as "not a migrant." Accompanying the screenshot is a short paragraph that explains that the species cannot take advantage of U.S. bird populations because they do not migrate and explicitly informs participants that the claim is false.

The third part of our experiment is identical for all treatments. Similar to Experiment 1, we elicit four incentivized outcome measures. Subjects have the opportunity to donate up to $1 to the Galapagos Conservation Trust (GCT), a charity that focus exclusively on the conservation and sustainability of the Galapagos Archipelago, to help the vampire finch.[29] The donation interface is exactly the same as in the Experiment 1. The second and the third incentivized measures (the

[29] We also include a small paragraph describing the species: "The Galápagos finch species collectively form a showcase example of Charles Darwin's theory of natural selection. The species of Galápagos finches are often called "Darwin's finches." They are used as an example of how the descendants of one ancestor can evolve through adaptive radiation into several species as they adapt to different conditions on various islands. Once we have completed our study, we will add all the donations together and send them to Galapagos Conservation Trust and submit the donations as an anonymous donor. At the end of this study, we will also provide you with a link to the website of Galapagos Conservation Trust so that you can learn more about them."

descriptive norm of beliefs about donations from our participants and the prescriptive norm regarding the appropriateness of different donation amounts) and their corresponding interfaces remain exactly the same as in the Migrant Caravan Experiment.

In our fourth incentivized measure, we ask participants to indicate the number of vampire birds that migrated northward and crossed the U.S. border in 2016. Again, in the debunking conditions, this allows us to test if debunking leads to a more accurate belief about the vampire birds. A participant receives an additional $0.25 payment for the correct answer (which is "zero").

Finally, the following contextual adjustments are made to the questionnaire in Experiment 1 to fit the context in this experiment. We elicit personal feelings about migratory birds that come to the U.S. ("positive", "negative" or "no strong feelings"), and pre-experiment knowledge about the species. Participants in the treatment conditions (false-claim, debunk-before, and debunk-after) are also asked about the extent to which exposure to the claim affected their donation decision as well as any pre-experiment exposure to species and the claim.

Our total sample consists of 453 participants who are assigned to one of six sessions, with 132 participants in the Baseline condition, 102 in False-claim, 115 in Debunk-before, and 104 in Debunk-after. The recruiting procedures and criteria are exactly the same as Experiment 1. Table 6 shows that our randomization procedure is successful.

Table 6: Balance of Covariates

	Baseline	False-claim	Debunk-before	Debunk-after	p-value
	(1)	(2)	(3)	(4)	(5)
Female	0.417	0.353	0.365	0.423	0.626
	(0.495)	(0.480)	(0.484)	(0.496)	
Education	4.265	4.186	4.174	4.115	0.417
	(0.640)	(0.656)	(0.639)	(0.828)	
Age	38.89	36.36	37.37	36.89	0.272
	(10.53)	(10.74)	(10.07)	(10.36)	
Republican	0.273	0.235	0.270	0.240	0.881
	(0.447)	(0.426)	(0.446)	(0.429)	
Heard Species	0.197	0.157	0.165	0.221	0.604
	(0.399)	(0.365)	(0.373)	(0.417)	
N	132	102	115	104	

Notes: Columns (1) to (4) report the mean level of each variable for each of our four conditions, with standard deviations in parentheses. Column (5) indicates the p-value when testing that means are the same across conditions. The variables "Female," "Republican," and "Heard Species" are dummy variables. "Heard Species" equals one if for subjects with pre-experiment exposure to the information about vampire ground finch. Education is an ordinal variable coded with following specifications: 1: Elementary school, 2: Junior high school, 3: Senior high school, 4: Undergraduate school (College), and 5: Graduate school (Masters or professional).

3B. Results

Result 1 is qualitatively and quantitatively replicated in this context, with a higher statistical significance. From the first two shaded bars in Figure 2, we see that the average donation in the false-claim condition is significantly lower than that in the baseline condition.

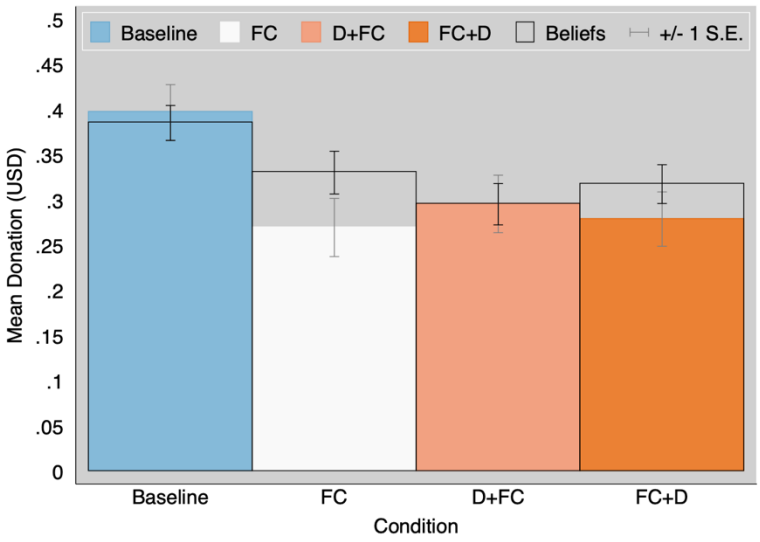

Figure 2: Mean Levels of Donations (shaded bar) and Mean Beliefs Regarding Other's Donations (framed bar) out of 1 USD.

Individuals exposed to a false claim significantly decreases the average donation toward migrants by 32% (the same percentage as in the first experiment), from \$0.397 of the available dollar to \$0.270 (p-value=0.003, one-tailed t-test). Again, the results from a Wilcoxon-Mann-Whitney test reinforce this finding as they allow us to reject the conjecture that the False-claim donation distribution is equal to that of the Baseline (p-value<0.01). Hence, we conclude that exposure to a false claim regarding the migratory birds also negatively impacts participants' decisions to make a donation to an organization supporting the species.

Result 1B (Donations): *The average donation in the false-claim condition is 32% lower than the average donation in the baseline condition (p<0.01).*

We now test whether the mechanism by which false claims operate is to reinforce an incorrect prior belief using the same hurdle model in Experiment 1. [30]

Table 7: Hurdle Model Estimation on Donation

	Full sample	**Only those with incorrect priors**
False-claim: hurdle	-0.481**	-0.682***
	(-2.82)	(-3.54)
False-claim	-0.075	-0.051
	(-1.11)	(-0.72)
Average marginal effect	-0.127**	-0.155**
	(-2.88)	(-3.13)
N	234	194

Notes: The z-statistics are reported in parentheses. The average marginal effect of the false claim on donation is reported for each model as the coefficient estimates are not directly interpretable. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Column one in Table 7 shows the results when we include all subjects (those with correct and incorrect priors). We find directional evidence that exposure to the false claim reduces the donation probability relative to the baseline and this result is statistically significant (z=-2.82, p-value<0.01). Conditional on donating, there is no difference in the amount donated (z=-1.11, n.s., column 1).[31] Column two runs the same regression, but only for those subjects who have incorrect priors. We see that the donation probability is lower (z=-3.54, p-value<0.001) and, conditional on donating, there is no difference in the amount donated (z=-0.72, n.s., column 2). These results replicate what we find in Experiment 1. Hence, the result that the effect of false claims on behavior happens mainly at the extensive margin seems robust across different contexts. This result supports

[30] A correct prior belief means that a subject answers "zero" to the question "guess the number of vampire birds that migrated northward, crossing the U.S. border, in 2016." Hence, choosing a non-zero answer indicates that the subject has an incorrect prior.

[31] Although our coefficient estimates are not directly interpretable, a post estimation shows that exposure to the false claim lowers the donation by about $0.127 (z =-2.88, p-value<0.01), which is 32% of the baseline donation level. Wilcoxon-Mann-Whitney tests also show similar results regarding the comparison between donation distributions in the false-claim and baseline conditions. Specifically, these results allow us to reject the conjecture that the baseline and false-claim donations are from the same distribution (z=2.93). However, the donation distribution is not significantly different when it is conditional on one being a donor (z=1.03). These results fully replicate what we found in Experiment 1.

hypothesis 2a and 2b that exposure to the false claim reinforces an incorrect prior belief such that subjects in the false-claim condition are less likely to make a donation.

> ***Result 2B (Mechanism: Beliefs reinforced):*** *The hurdle model finds evidence consistent with reinforcement models: A false claim lowers the probability of donating in the false claim treatment (relative to the baseline). This effect is pronounced among those with incorrect priors ($p<0.01$). However, conditional on donating, there is no difference in the amount donated. Taken together, these results support a mechanism whereby false claims are reinforcing an incorrect prior belief.*

We now turn to testing hypotheses 3a-3e regarding the impact of exposure to a false claim on what people consider appropriate to do— i.e. testing if the norm is changed. Table 8 presents regressions testing the first three hypotheses.

Different from Experiment 1, the false claim in Experiment 2 marginally changes the descriptive norm but not the prescriptive norm. Column one finds marginally significant differences between treatments regarding beliefs about what others will donate (beta=-0.051, $p<0.1$). On average, individuals believe others will donate \$0.385 in the baseline condition, but only \$0.330 in the false-claim condition. However, column two shows that exposure to a false claim does not change the social appropriateness of donating zero (beta=0.087, n.s.).

Because the descriptive norm of donating has changed in the vampire bird scenario, hypothesis 3c tests whether this is correlated with an increase in zero donations. Column three finds support for this: There are more zero donations after exposure to the false-claim. The result is significant at 1% level after controlling for prior belief, education, age, gender and political ideology.

Table 8: The Effect of Exposure to the False Claim on Social Norms and Donation

DV	(1) OLS Descriptive norm ratings	(2) OLS Prescriptive norm for donating zero	(3) LOGIT Donate zero=1 Marg. Effects
False claim	-0.051^{+}	0.087	0.162**
	(-1.82)	(0.62)	(2.95)
+Controls	Yes	Yes	Yes

Constant	0.031	1.636**	
	(0.28)	(2.98)	
N	234	234	234

Notes: The control variables include education, age, gender, political orientation and the prior belief. The *t*-statistics are reported in parentheses; marginal effects are reported for logit regressions. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Thus, we find that in the migrant caravan scenario, a false claim changed the prescriptive norm and made donating zero more appropriate. However, in the vampire bird scenario, the false claim changed participants' descriptive norm and caused them to believe that others were likely donating less. In both scenarios, there were more zero donations after exposure to the false-claim.

Following the line of analysis for experiment 1, the conditional logit estimates in Table 9 present evidence that the descriptive norm predicts donations (beta=0.063, p-value<0.001, col. 2) while the interaction term shows that the weight a participant attaches to a prescriptive norm does not differ by treatment (beta=-0.017, n.s., col. 3). In addition, the comparison between the BIC (Bayesian information criterion, which penalizes model complexity) of models one and two shows that including the descriptive norm rating when predicting the chosen action of donation increases the model's predictive fit.

Result 3B (Mechanism: Descriptive norm change): *In comparison with the baseline, false-claim exposure causes the descriptive norm to weaken, leading to belief that others will donate less, and causes an increase in zero donation amounts. Regression estimates suggest this behavioral change is not driven by changes in the weight placed on norm following between baseline and false claim.*

Table 9: Conditional Logit on the Descriptive Norm and Chosen Donation Amount

	(1)	(2)	(3)
	chosenAction	chosenAction	chosenAction
Payoff to Self	6.262***	3.250	3.171
	(3.71)	(1.45)	(1.43)
Belief of others' donation		0.063***	0.073***
		(10.61)	(8.28)
Belief of others' donation x False Claim			-0.017
			(-1.60)
+Controls	Yes	Yes	Yes
AIC / BIC	968.3./1266.8	768.4/1072.8	767.8/1078.0

N	2574	2574	2574

Notes: The control variables include education, age, gender, political orientation and the prior belief of DACA eligibility. *t* statistics in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Results two and three test for two channels through which a false claim impacts behavior: It can reinforce incorrect priors and change the amount subjects believe that others will donate (the descriptive norm of donation). We now investigate our debunking treatment for further evidence in support of these channels.

We first show that subjects with inaccurate priors update beliefs in our debunking conditions. We find that 21.74% (35.58%) correct their beliefs and realize that vampire birds do not migrate northward to the U.S. in the before (after) debunking treatment, and the proportion of subjects updating is significantly different from that of the false-claim and the baseline condition (two-sample test of proportions p-value<0.001 for all four pairs of comparisons, z=-4.41, -6.16, -5.08., -6.99).[32] Note that we do not find a difference in the proportion with correct prior beliefs in debunking conditions from conditions without debunking (z=0.59).

> ***Result 4a (Mechanism: Beliefs changed):*** *Exposure to debunking information leads to more accurate beliefs compared to those who receive a false claim without debunking information.*

However, consistent with our finding in Experiment 1, debunking is again only partially effective in resorting donation behavior to baseline levels by increasing donations. The average donation after debunking ($0.30 and $0.28 for debunking-before and debunking-after, respectively) hovers between levels observed in the baseline and false claim condition. In this experiment, it is not significantly higher than the mean donation in the False-claim condition ($0.27) but significantly lower than the baseline ($0.40). When we only consider those who update in the debunk before/after treatment (22% and 36%, respectively), surprisingly, we find that these participants donate significantly less than those who do not update their beliefs (p-value<0.001 in both

[32] When consider the timing of debunking, the proportion of subjects who correct update their beliefs in the debunk-before treatment is significantly different from (lower) that in the debunk-after at 5% level (z=-2.26) in this experiment.

debunking conditions). Lastly, we find that subjects who accurately update their beliefs after debunking donate significantly less than participants in our false-claim condition (p-value<0.05 in the debunk-before and p-value<0.01 in the debunk-after condition).

> ***Result 4b (Donation: Beliefs changed)****: The average donation of those in debunking conditions is not significantly more than the false claim condition and significantly lower than the baseline.*
>
> ***Result 4c (Debunking: Beliefs reinforced)****: Conditional on having incorrect priors, the average donations of those update to the correct post belief after debunking is significantly less than those retain their incorrect priors.*

When we examine the impact of debunking on norms, we find evidence that supports the interpretation that false claims change norms. Consistent with results 3a-3c in experiment 2, in the debunking treatments, subjects still believe others will donate less ($0.30/0.32) compared to those in the baseline ($0.39) (p-value<0.01, p-value<0.05, two-tailed). [33] The difference is even larger when subjects correct their false priors. They still believe others will donate less ($0.19) compared to those in the baseline ($0.39). This difference in the descriptive norm is significant (p-value<0.001, two-tailed).[34]This result also holds when we limit the baseline sample to individuals with incorrect priors (p-value<0.001, two-tailed). Thus, even successful debunking cannot reverse the damaging impact of a false claim on norms to donate.

> ***Result 4d (Debunking: Fail to recover descriptive norms)****: Individuals who experienced successful debunking in the debunking conditions believe others will donate less compared to those in the baseline / with incorrect priors. Thus, again, hypothesis 4d is not supported.*

In aggregate, results 4a-4d support the interpretation that false-claim exposure causes changes in behavior by changing the descriptive norms. This effect is irreversible even after successful debunking. Although exposure to a false-claim seems to reinforce an incorrect prior, correcting the

[33] The average belief about others' donation is 0.30 (0.32) in debunk-before (-after) and the pooled average belief about others' donation across the two debunking conditions is 0.31.

[34] Consistent with result 3b (false claim does not change prescriptive norm from the baseline), the average prescriptive norms rating of subjects in the debunking conditions who updated their incorrect prior to a correct posterior is 2.08, not significantly from the baseline.

false belief does not alleviate the negative impact of false-claim on behavior. Consistent with results from Experiment 1, overall, results from Experiment 2 provide stronger evidence that debunking does not fully mitigate the impact of a false-claim on behavior.

4. Discussion

Using two experiments on Amazon's Mechanical Turk, we find that exposure to a false negative claim regarding a subject (2018 migrant caravan or an endangered species) lowers the average donation towards it by 32%. We show that the false claim reinforces incorrect prior beliefs and changes what individuals consider the donation social norm to be. Although exposure to a false-claim seems to reinforce an incorrect prior, successfully correcting the belief by providing debunking information does not restore either contribution behavior or norms to baseline levels. However, we also document that the channel through which false claims impact behavior is different for the migrant caravan (where it changes beliefs about the injunctive norm) and the vampire birds (where it changes beliefs about the descriptive norm).

One possibility for why false claims affect different beliefs about norms is that subjects likely had little prior knowledge or attitudes about vampire migrant birds. Indeed, the "threat" was something that we created for this study and, unlike the migrant caravan, had no prior news coverage. As such, subjects likely had little to no prior exposure to issues related to vampire birds. Though purely speculative, it is possible that the vampire bird treatment may have led to greater cognitive deliberation while the migrant caravan activated implicit attitudes and activated more automatic processes (Farrow et al. 2017; Bogo et al. 2020). Other research indicates that when individuals are under a cognitive load, the influence of descriptive norms on behavior increases while the influence of injunctive norms decreases (Melnyk et al., 2011). Melynk et al. (2011) provide support for the hypothesis that compliance to descriptive norms can constitute a heuristic shortcut that reduces the effort involved in decision-making when cognitive resources are limited.

By contrast, it is possible that in the migrant caravan treatment, the effects are mediated by activating existing underlying attitudes toward migrants (positive, neutral, or negative). We can test this with data we collected that comes from a question asked at the end of our experiment. The question was "Do you have personal feelings about immigrants to the US?" The response format was "I have positive/ no strong/ negative/ feelings about immigrants to the US."

We estimate an OLS model with main effects for the false-claim exposure, attitudes toward migrants and an interaction of false-claim exposure and attitudes (we also include control variables).[35] We present the regression in the appendix and focus our discussion on a graphical representation of the regression. Figure 3 shows the predicted relationship between the norms rating and attitudes by the baseline and false-claim condition (with 95% confidence intervals). On the y-axis are the predicted averages for the norm rating to donate zero. The norm ratings range from "very socially inappropriate=1," "somewhat socially inappropriate=2," "somewhat socially appropriate=3," or "very socially appropriate=4".

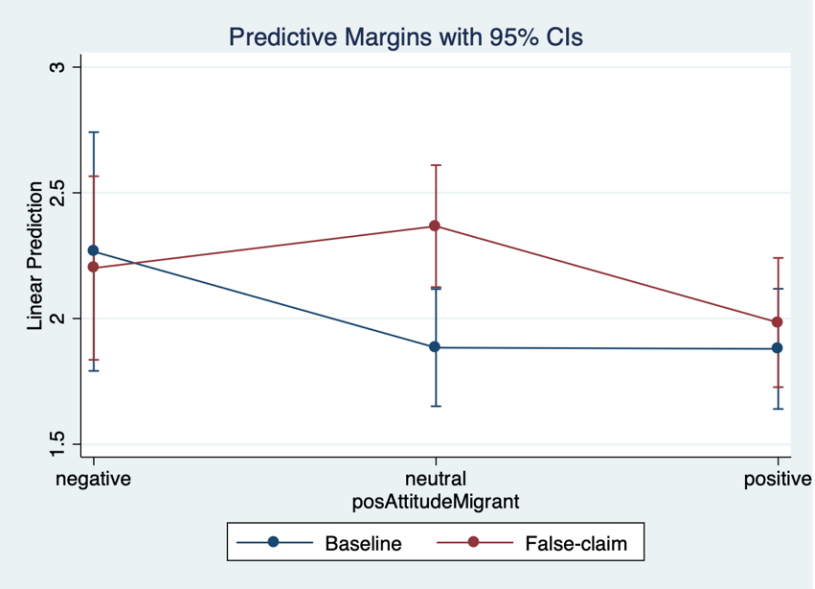

Figure 3: Average Predicted Norm Rating of Donating Zero
Notes. "very socially inappropriate=1," "somewhat socially inappropriate=2," "somewhat socially appropriate=3," or "very socially appropriate=4."

We see that having a positive attitude toward migrants is correlated with a lower predicted norm rating. That is, respondents who have a positive attitude towards migrants, also believe that donating zero is inappropriate (average ratings in the baseline=1.879 and false-claim=1.984, difference is n.s.). Having a negative attitude towards migrants is correlated with believing that it is appropriate to donate zero to migrants (average appropriateness ratings in the baseline=2.266 and

[35] Control variables include education, age, gender, political orientation and the prior belief of DACA eligibility. Detailed regression results are reported in Table A.1.

false-claim=2.201, difference is n.s.). However, exposure to the false-claim changes the normative beliefs of those who have a neutral attitude —they believe it is appropriate to donate zero in the false-claim condition (norm rating=2.367) but believe it is inappropriate in the baseline (norm rating=1.884). This difference is significant at the $p<0.01$ level.

This result provides evidence that misinformation affects the appropriateness of not donating in a manner that is different from confirmation bias and selective exposure (Chaffee and Miyo, 1983; Eliaz and Schotter, 2006; Nickerson, 1998; Karlsson, Lowenstein, and Seppi, 2005). Thus, we take these results as expanding our understanding of how misinformation may impact charitable giving—misinformation impacts the appropriateness of donating for those who do not yet have strong attitudes.

5. Conclusion

In this study, we use incentivized outcome measures to examine the effects of negative false claims and debunking information on behaviors and perceptions of social norms. Our results provide evidence for the persuasive effects of false claims on charitable decision making.

Using two experiments on Amazon's Mechanical Turk, we find that exposure to a false negative claim regarding a 2018 migrant caravan or an endangered vampire bird lowers the average donation towards a cause that would support the affected group by 32%. We show that the false claim reinforces incorrect prior beliefs and changes what individuals consider the donation social norm to be. Although exposure to a false-claim seems to reinforce an incorrect prior, successfully correcting the belief by providing debunking information does not restore either behavior or norms to baseline levels. Our findings indicate that disinformation turns out to be a sticky threat – once present, it is difficult to undo its harm on behavior or social norms.

References

Allenby, G., Belk, R., Eckel, C., Fisher, R., Haruvy, E., Ma, Y., ... and Li, S. X. (2020). *Fundraising Design: Key Issues, Unifying Framework, and Open Puzzles* (No. 00683). The Field Experiments Website.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, *100*(401), 464-477.

Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544-55.

Bicchieri, C. and Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public choice*, 1-22.

Bicchieri, C. and Mercier, H. (2014). Norms and beliefs: How change occurs. In *The complexity of social norms* (pp. 37-54). Springer, Cham.

Blough, D. S. and Millward, R. B., (1965). Learning: operant conditioning and verbal learning. *Annual Review of Psychology*, 16(1), 63-94.

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of experimental psychology: general*, *149*(8), 1608.

Chaffee, S. H., and Miyo, Y. (1983). Selective exposure and the reinforcement hypothesis: An intergenerational panel study of the 1980 presidential campaign. *Communication Research*, 10(1), 3-36.

Chambers, K. L., and Zaragoza, M. S. (2001). Intended and unintended effects of explicit warnings on eyewitness suggestibility: Evidence from source identification tests. *Memory & Cognition*, 29(8), 1120-1129.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, *58*(6), 1015.

Cragg, J. G. (1971). Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica: Journal of the Econometric Society, 829-844.*

DellaVigna, S. and Kaplan, E. (2007). The Fox News Effect: Media Bias and Voting. *Quarterly Journal of Economics*, 122(3), 1187-1234.

DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The quarterly journal of economics*, *127*(1), 1-56.

Deutsch, M., and Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology*, 51(3), 629.

Ecker, U. K., Hogan, J. L. and Lewandowsky, S. (2017). Reminders and Repetition of Misinformation: Helping or Hindering Its Retraction?. *Journal of Applied Research in Memory and Cognition*, 6(2), 185-192.

Ecker, U. K., Lewandowsky, S., and Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, 38(8), 1087-1100.
Farrow, K., Grolleau, G., & Ibanez, L. (2017). Social norms and pro-environmental behavior: A review of the evidence. *Ecological Economics*, *140*, 1-13.

International Broadcasting Trust (2018). *Faking It: Fake News and How it Impacts on the Charity Sector*. International Broadcasting Trust. https://www.ibt.org.uk/reports/faking-it/

Internal Revenue Service (2020). *IRS warns against COVID-19 fraud; other financial schemes*. Internal Revenue Service. https://www.irs.gov/newsroom/irs-warns-against-covid-19-fraud-other-financial-schemes

Jack, C. (2017). Lexicon of lies: Terms for problematic information. *Data & Society*, 3, 22.

Jou, J., and Foreman, J. (2007). Transfer of learning in avoiding false memory: The roles of warning, immediate feedback, and incentive. *Quarterly Journal of Experimental Psychology*, 60(6), 877-896.

Karlsson, N., Loewenstein, G., & Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and uncertainty*, 38(2), 95-115.

Kimbrough, E. O., and Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3), 608-638.

Krupka, E. and Weber, R. A. (2009). The Focusing and Informational Effects of Norms on Pro-social Behavior. *Journal of Economic Psychology*, 30(3), 307-320.

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

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger,M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J. and Zittrain, J. L. (2018). The science of fake news. *Science* 359(6), 1094-1096.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N. and Cook, J. (2012). Misinformation and Its Correction. *Psychological Science in the Public Interest*, 13(3), 106-131.

Martin, G. J. and Yurukoglu, A. (2017). Bias in Cable News: Persuasion and Polarization. *American Economic Review*, 107(9), 2565-2599.

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behavior.

Melnyk, V., Herpen, E. V., Fischer, A. R., & van Trijp, H. C. (2011). To think or not to think: The effect of cognitive deliberation on the influence of injunctive versus descriptive social norms. *Psychology & marketing*, *28*(7), 709-729.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175-220.

Pennycook, G., Cannon, T. D. and Rand, D. G. (2018). Prior Exposure Increases Perceived Accuracy of Fake News. *Journal of Experimental Psychology General*, 147(12), 1865-1880.

Ruths, D. (2019). The misinformation machine. *Science*, 363(6425), 348-348.

Tandoc Jr, E. C., Lim, Z. W., and Ling, R. (2018). Defining "fake news" A typology of scholarly definitions. *Digital journalism*, 6(2), 137-153.

Skinner, B. F. (1937). Two types of conditioned reflex: A reply to Konorski and Miller. *Journal of General Psychology*, 16(1), 272-279.