HATE, FEAR AND INTERGROUP CONFLICT

Experimental Evidence from Nigeria

Miguel Ortiz*
UC Berkeley
(Job Market Paper)

October 19, 2023

Click here for latest version

Abstract

Understanding the deep drivers of social conflict is crucial for identifying policies that can effectively reduce it. In this paper, I study to what extent intergroup conflict is driven by hate (a preference for harming the outgroup) vs. fear (a belief about the outgroup's hate towards the ingroup), and how policy interventions affect these drivers to increase cooperation. To this end, I develop a theory-driven experimental protocol to disentangle these two motives and determine their relative importance empirically. I deploy this protocol as a lab-in-the-field experiment in Jos, Nigeria, to study the region's ongoing conflict between Christians and Muslims. I find that fear explains 76%, and hate 24%, of the non-cooperative behavior I observe in a coordination game played between Christians and Muslims. Moreover, this fear is mostly unwarranted, as non-cooperators grossly exaggerate the percentage of hateful people in the other group. I then estimate a structural model to determine what type of policy intervention would most effectively increase cooperation. My counterfactual analysis suggests that interventions that correct unwarranted fears would be highly effective. In contrast, interventions that reduce hate would not because hateful people also have high levels of fear. Finally, I study an actual policy intervention with an RCT in which I provide participants access to a radio drama that promotes intergroup cooperation. Using my experimental protocol, I find that the radio drama decreases hate but not fear and thus does not translate into increased cooperation, as my model predicts.

*m.ortiz@berkeley.edu. I am especially grateful to Ernesto Dal Bó and Francesco Trebbi for their generous support and guidance throughout these years. I am also grateful to Ned Augenblick, Matilde Bombardini, Fred Finan, Sam Kapon, Edward Miguel, Ryan Oprea, Ricardo Pérez-Truglia, Jonathan Weigel and Guo Xu for their comments and support. Eunice Boba Atajiri did an outstanding job as the field manager of the experiments. I gratefully acknowledge the funding provided by the Association for Comparative Economic Studies, the Center for Effective Global Action, the Center for Equity, Gender & Leadership, and the Institute for Business Innovation. The pre-analysis plan of this paper is available at *socialscienceregistry.org/trials/10605*.

1. Introduction

Societies are often fractured into social identity groups, like ethnicities or religions, which results in significant economic losses (Easterly and Levine, 1997; Alesina and La Ferrara, 2005) and paves the way for the rise of extremist leaders and social conflict[1]. Extremist leaders commonly believe that the prosperity of their own group (the ingroup) hinges on taking hostile actions against the other group (the outgroup). Those hostile actions, and the resulting social conflict, can take many forms, encompassing labor discrimination and the curtailment of civil rights and access to public goods, as well as segregation, oppression and, in the gravest circumstances, war and genocide. Despite this being a recurrent pattern in history and across the globe, we still do not have clarity on what motive drives citizens to support these hostile actions against the outgroup and fuel social conflict. Support for such hostilities may be rooted in preferences: individuals harbor hatred towards the outgroup and want to harm them. Alternatively, this support could stem from beliefs: individuals do not hate the outgroup but are afraid that the outgroup wants to harm them, causing them to support hostile actions to protect themselves. This fundamental distinction is of high importance, as different drivers of conflict will call for different types policy interventions to promote intergroup cooperation. Nevertheless, empirically disentangling the underlying drivers is hard.

In this paper I seek to answer two questions. (1) To what extent is intergroup conflict driven by hate vs. fear? I define hate as a *preference* for harming the outgroup, and fear as a *belief* about the outgroup's hate towards the ingroup. I conceptualize social conflict as a manifestation of non-cooperation, in the tradition of Fearon and Laitin (1996). Here, I further ask: If conflict is driven by fear, are beliefs accurate or misperceptions? After understanding what drives conflict, I turn to policy and ask: (2) Are policy interventions currently trusted to promote cooperation the right policies for the task? By which I mean, are these interventions able to address what I identified to be the key driver of conflict? Here, I focus on the creation of media content to promote intergroup cooperation—in particular, radio drama series.

The analysis takes place in Jos, a state capital situated in the region of Nigeria where the Muslim North and Christian South meet. Historically, the city has been inhabited by both religious groups, co-existing peacefully. However, with democratization in the 1990s, political leaders with religious banners began fighting for power, leading the city to increased tensions and the threat of violence. In the 2000s, the city experienced multiple outbreaks of religious violence perpetrated by ordinary citizens from both sides. These events created mistrust and animosity between religious groups and led to a process of segregation in all aspects of life. Today, there is little interaction between the groups, religion is the key political cleavage, and politicians fuel negative narratives about the outgroup for political gain.

To answer question (1), I develop a theory-driven experimental protocol to empirically disentangle

[1]The literature on this is ample. Regarding the rise of extremist leaders see, for example, Guriev and Papaioannou (2022), Colussi et al. (2021), Brunner and Kuhn (2018), Dustmann et al. (2019). And regarding the rise of intergroup conflict see, for example, Esteban, Mayoral and Ray (2012), Cederman et al. (2009).

the motives driving intergroup conflict and asses their relative importance. Then, I deploy this protocol as a lab-in-the-field experiment, in Nigeria, to study the conflict between Christians and Muslims in this country. I start by writing down a model of conflict, with hate and fear as primitives. I model conflict between groups as a coordination game, where cooperation is an equilibrium and offers the highest possible payoff to each player. In this game, players may prefer to not cooperate if they feel enough hate for the outgroup—that is, if a player prefers to sacrifice part of her payoff in order to reduce in a greater amount the payoff of the outgroup player. Alternatively, non-cooperation may stem also from fear. A player who is not hateful but fears the outgroup player is hateful (and therefore non-cooperative) will also want to not cooperate.

In the field, I measure cooperation between Christians and Muslims through coordination games. However, as noted before, not cooperating is an equilibrium outcome that is driven by both preferences and beliefs. To disentangle these drivers, I use the following insight when designing the experimental protocol. For each coordination game that subjects could play, it is possible to design a money allocation decision that mirrors the structure of the coordination game but removes the uncertainty in payoffs, such that beliefs do not enter the empirical problem, only preferences. In this way, the money allocation decision isolates the preferences that play a role in the decision in the coordination game. With this in mind, in the experiment, participants make a series of money allocation decisions that elicit their willingness to pay to decrease or increase the payoff of an outgroup member. This allows me to recover their level of hate (or altruism) in a way that is directly connected to their decision in the coordination game. To elicit fear, I ask participants to guess the money allocation decisions of other participants to elicit their beliefs about outgroup members' level of hate (or altruism). I then use all the information to estimate a structural model to properly recover social preferences at the individual level. Using the estimated preferences and elicited beliefs, I determine the extent to which non-cooperation is driven by hate vs. fear. I then use the estimated model to conduct counterfactual analysis and study how policy interventions that shift hate or fear would affect cooperation. In particular, I show how cooperation would change if a hypothetical intervention were to (i) solve unwarranted fears by correcting misperceptions about the outgroup, or (ii) completely eradicate intergroup hate.

The first result is that in a game where cooperating is both an equilibrium and the maximum payoff (which represents half a day of salary in this context), people fail to cooperate in 31% of the interactions between groups—compared to only 6% of interactions within groups. This leads to a loss of 9.6% of attainable wealth in intergroup interactions. I estimate the model and find that it performs well in terms of sample fit: the hate and fear elicited from participants explain over 90% of the decisions made in the game. The estimated model leads to three main findings. First, 24% of non-cooperation decisions were motivated by hate, while 76% were motivated by fear. Second, fear is mostly unwarranted, as non-cooperators grossly exaggerate the percentage of hateful people in the outgroup. Third, hateful individuals tend to also be very fearful of the outgroup, while altruistic individuals show a wide range of beliefs, from fearful to trusting. The counterfactual analysis reveals that if a policy solved unwarranted

fears by correcting inaccurate beliefs about the outgroup, the number of people not cooperating would drop by 73%. This result underlines how misperceptions leading to unwarranted fears are the most important barrier to intergroup cooperation. In contrast, if a policy completely eradicated intergroup hate, the number of people not cooperating would drop by only 5%. The effect on cooperation is small in this case because hateful individuals are also very fearful and therefore, even without hate, most of them will still want to not cooperate out of fear. Importantly, I show that my measures of hate and fear correlate with support for policies of religious segregation—and, what is more, my measure of hate correlates more strongly with support for segregation for hateful reasons, while my measure of fear correlates more strongly with support for segregation for fearful reasons.

Furthermore, I find important differences in the way Christians and Muslims behave. Specifically, 84% of the decisions to not cooperate come from Christians and 16% come from Muslims. Additionally, Christians have more negative social preferences towards Muslims and more biased beliefs about them than the other way around. These results are in line with what was predicted in the pre-analysis plan, as I expected heterogeneity in this direction based on extensive fieldwork. The difference in behavior between religious groups is most likely driven by the salience of the armed group Boko-Haram in this area, which generates negative feelings towards only one of the groups.

Having understood what drives intergroup conflict, I turn to question (2), which asks if popular policy interventions are effective at addressing the key driver of conflict and promoting cooperation. To this end, I conduct a randomized control trial (RCT) where I randomly give participants access to a radio drama series that promotes intergroup cooperation, and evaluate its effects on hate and fear. Radio dramas are both a popular form of entertainment in Sub-Saharan Africa and a common intervention used by NGOs in the region. Recently, they have received increased attention as a policy for conflict due to their perceived advantages: fictional stories make it easier to address sensitive topics of conflict (Slater and Rouner, 2002); stories increase attention and retention of the message (Kromka and Goodboy, 2019); and media interventions can be implemented in a wide range of contexts where alternative policies for conflict are unfeasible. To study this policy, I partnered with the radio production company hired by the largest NGOs in Nigeria, and produced a new radio show following the exact same steps NGOs take to produce their shows. This new radio drama aimed to reduce both hate and fear: the story is about two communities that, driven by hate and unfounded fears, miss out on mutually beneficial interactions, and its resolution has a message on letting go of hate and reevaluating fear. The treatment consisted of 24 episodes, each lasting between 10-15 minutes. The show was not broadcasted, but instead episodes were sent to participants through WhatsApp, four times a week over a six-week period. To promote and monitor engagement, participants were incentivized to answer weekly quizzes on the show's content. The control group listened to a placebo radio drama with a message on health. At endline, participants went through the lab-in-the-field experiment again, which allowed me to examine if hate and/or fear were impacted and how this affected cooperation.

I find that the radio show treatment is effective at reducing hate (by 0.45 SD), but ineffective at

reducing unwarranted fears. Furthermore, the treatment proves ineffective at increasing cooperation. The model allows to rationalize what could otherwise be a puzzling result. The radio show is an effective policy because it reduces hate, but it is the wrong policy for this context because it does not affect the key motive of non-cooperation, which is fear. This ultimately renders the policy ineffective at achieving its main goal, which is increasing cooperation. Beyond this, I find that the effect on preferences is strongest in the most hateful subsample—a result that was not obvious *a priori*, as the most hateful individuals could have had more rigid preferences. I also find some evidence that the radio show did reduce the fear of those who had the most biased beliefs. Importantly, I show the results are not likely driven by social desirability bias, following the methodology in Dhar et al. (2022).

Taken together, the results of this paper illustrate the value of the model-based protocol to explain conflict behavior and policy efficacy. The use of the protocol is twofold. First, the protocol serves to diagnose what drives conflict in a particular setting. Second, it serves to determine which drivers of conflict a specific policy can shift. Furthermore, by comparing these two results, the protocol serves to assess the alignment or misalignment between a place's needs and a policy. For the case of this paper, the two results reveal a mismatch: a radio show that decreases hate will not increase cooperation in Jos, where the main driver of conflict is fear. Indeed, the null treatment effect I find on cooperation confirms this misalignment. Importantly, this protocol is portable and can be deployed elsewhere. Using it for other settings and policies can advance the understanding of intergroup conflict and its solutions.

This paper makes three main contributions. First, it presents a novel theory-driven experimental protocol to empirically disentangle hate and fear in a way that proves useful at explaining conflict behavior and policy efficacy. The literature has provided evidence of how group membership affects preferences and beliefs (for reviews see Shayo (2020) and Charness and Chen (2020)). Concerning preferences, group membership has been shown to affect social preferences positively and negatively (Chen and Li, 2009; Choi and Bowles, 2007; Fershtman and Gneezy, 2001; Bauer et al., 2014; Kranton et al., 2020; Enke et al., 2023). Concerning beliefs, group membership has been shown to affect trust, stereotypes and prejudice (Bénabou and Tirole, 2011; Falk and Zehnder, 2013; Bonomi et al., 2021). This papers builds upon this literature by using its finding to develop a theory of intergroup conflict, and its tools to develop a protocol that estimates the parameters of the theory.

Second, this paper provides empirical evidence on how unwarranted fears can be an essential driver of conflict and how policies may struggle to solve them. This finding directly contributes to the theoretical literature on fear and conflict, (Chassang and Padró i Miquel, 2007; Acemoglu and Wolitzky, 2023), which lacks empirical evidence. In addition, the results show how in a context that can be theoretically described by "the politics of fear" (Padró i Miquel, 2007), fear indeed becomes a core problem of intergroup relations, and what is more, one that proves hard to fix with policy. More generally, the finding contributes to the literature studying the dynamics of conflict and intergroup cooperation (Fearon and Laitin, 1996; Blattman and Miguel, 2010; Bauer et al., 2016; Trebbi and Weese, 2019). Furthermore, the findings help explain what drives the inefficiencies that appear in all sorts of intergroup interac-

tions: for example, in labor selection (Oh, 2023; Giuliano et al., 2009), performance (Hjort, 2014; Ghosh, 2022; Marx et al., 2021; Alesina and Ferrara, 2005), trade (Korovkin and Makarin, 2023; Anderson, 2011; Michelitch, 2015; Jha, 2013), public spending (Luttmer, 2001; Franck and Rainer, 2012; Hodler and Raschky, 2014; Francois et al., 2015; Kramon and Posner, 2016), and political accountability (Casey, 2015).

Third, this paper provides experimental evidence on the effectiveness of using narrative media content as an intervention to foster intergroup cooperation. This evidence contributes to the literature on policies to improve cooperation between groups in conflict (for a review see Paluck et al. (2021)). Some of the interventions that have attracted academic attention lately are intergroup contact (Lowe, 2021; Rao, 2019; Mousa, 2020; Scacco and Warren, 2018; Paluck et al., 2019; Enos, 2014; Bursztyn et al., 2021; Fouka and Tabellini, 2022), perspective taking (Alan et al., 2021; Adida et al., 2018) and the use of narratives (Broockman and Kalla, 2016). The existing literature, however, lacks clarity regarding the mechanisms through which these interventions operate (Paluck et al., 2021). This paper sheds light on these mechanisms. In addition, this finding relate to the research on media and its effects on social and economic outcomes (for reviews see DellaVigna and La Ferrara (2015) and La Ferrara (2016)). More specifically, it contributes to the work studying how radio influences attitudes in conflict (Yanagizawa-Drott, 2014; Adena et al., 2015; DellaVigna et al., 2014; Paluck, 2009).

The rest of the paper proceeds as follows. Section 2 lays out the background of Jos, Nigeria. Section 3 presents the theory and experimental protocol. Section 4 describes the empirical model and estimation strategy. Section 5 describes the RCT on the radio drama and the data collection. Section 6 reports the results. Section 7 concludes.

2. Background

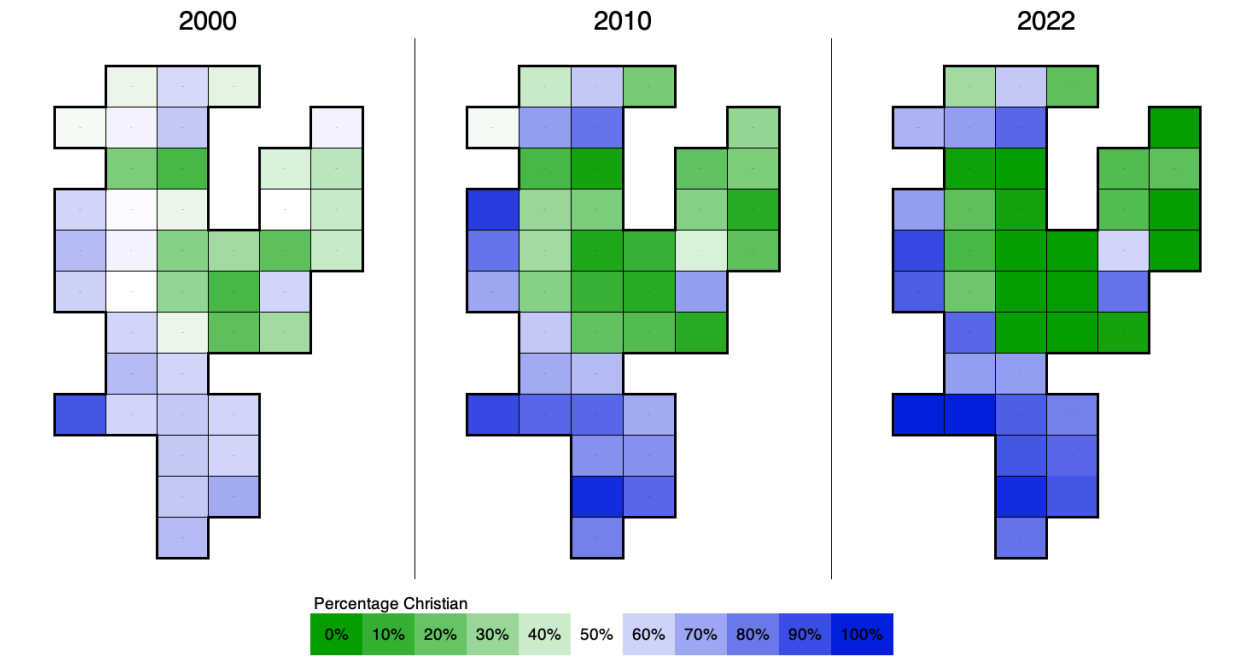
Nigeria is divided between a Muslim-dominated North and a Christian-dominated South. The Middle Belt is the region where these two religious communities intersect. Plateau is one of the states located in the Middle Belt and stands out as the most ethnically diverse state in the country. The experiments of this project took place in Jos, the capital of Plateau.

The city of Jos has historically had a balanced population of both Christians and Muslims. Throughout much of Jos' history, the coexistence between these two groups was characterized by peaceful and harmonious interactions. For instance, just 25 years ago both religious groups used to celebrate religious festivals together. However, with the onset of democratization in the 1990s, political leaders with religious banners began competing for power, leading to heightened tensions and a looming threat of violence. In 2001 occurred what came to be known as the 'First Crisis'. This was a spontaneous outbreak of inter-religious violence perpetrated by ordinary citizens, that spread throughout the city. The crisis lasted for several days and resulted in over a thousand fatalities, with both sides being both victims and perpetrators.

After the First Crisis, tensions between the religious groups increased even more This led to similar spontaneous outbreaks of inter-religious violence in 2004, 2008 and 2011, each resulting in hundreds of fatalities. The crises deeply scarred the city and broke what was left of the once-harmonious relationship between Christians and Muslims. Since the First Crisis, and reinforced by the ones that followed, a process of religious segregation was set in motion, permeating all aspects of life, such as residential areas, schools, jobs, local markets, hospitals and politics.

Figure 1 showcases the religious segregation that took place in Jos between 2000 and 2022. The figure shows a diagram for the map of Jos in three key points in time: 2000, just before the First Crisis; 2010, just before the 2011 crisis; and 2022, when the baseline data was collected. In each diagram, each quadrant represents one of Jos' communities (or neighborhoods) in its approximate geographical location. Green quadrants have a majority Muslim population, with darker greens representing a higher share of Muslims, and blue quadrants have a majority Christian population, with darker blues representing a higher share of Christians.

Figure 1: Religious Segregation in Jos, 2000-2022



Notes: The figure shows a diagram for the map of Jos at three points in time. In each diagram, each quadrant represents one of Jos' communities (or neighborhoods) in its approximate geographical location. Green quadrants have a majority Muslim population, with darker greens representing a higher share of Muslims, and blue quadrants have a majority Christian population, with darker blues representing a higher share of Christians. The data for this figure was collected as follows: Within the group participants of this project that were 40 or older and had lived in their community for at least 25 years, we picked 3 per community and asked them about the religious distribution of the population in these three years. For each community-year, I averaged the three answers and use that average as the data point.

Today, due to the segregation that took place, there is minimal interaction between the religious groups in the city. Although the city has regained some stability and safety over the past decade, the traumatic experiences of the past have made people reluctant to venture outside the areas of their re-

ligious group. This lack of contact has exacerbated mistrust and animosity between the groups. Additionally, religion has become the key political cleavage, with political parties using religious banners to mobilize voters in the quest for control over the city. Both sides fear that if the other religious group gains too much power, they may force them out of the city or block their growth in it. Many politicians exploit these fears for political gains, further exacerbating tensions. This context closely resembles the theoretical characterization of "the politics of fear", described by Padró i Miquel (2007).

Currently, power over the city is relatively balanced between the two religious groups, which may explain the fragile peace the city has experienced in recent years. However, this equilibrium is constantly under threat.

3. Framework and Experimental Protocol

This section presents the theory that guides the design of the experimental protocol and then proceeds to describe the lab-in-the-field experiment.

3.1. Framework

Consider a society with two groups, A and B. Let i be a member of group A, and j a member of group B. When interacting with j, i has the the following utility function:

$$u_i = x_i + \beta_i(z_i) \cdot x_j \tag{1}$$

Where x_i is i's payoff, x_j is j's payoff, and $\beta_i \in [-1, 1]$ is i's parameter of social preferences towards members of group B. If $\beta_i<0$, i is hateful towards members of group B; if $\beta_i>0$, i is altruistic towards members of group B; if $\beta_i=0$, i is selfish when interacting with members of group B. The bounds assumed on β_i signify that i can not care about j more that she cares about herself. z_i can be past experiences, education, etc. In what follows, I take z_i as given, but an alternative model can be found in the Appendix, where β_i is endogenous and depends on the beliefs about β_j.

Members of the different groups face each other in a coordination game[2] (or stag-hunt game), where there are two strategies: *Cooperate* (C) or *Not Cooperate* (N). I illustrate the theory with the following coordination game, which is the one subjects face in the experiment (where payoff units were in Nigerian naira). Payoffs in the matrix represent x_i and x_j in the utility function.

[2]Coordination games can allow for a richer study of the reasons behind non-cooperation, compared to other games like the prisoner's dilemma. In a prisoner's dilemma, non-cooperation is driven by selfish preferences ($\beta_i=0$). Instead, in a coordination game, selfish individuals could rationally want to cooperate. Therefore, non-cooperation in coordination games is driven by reasons beyond selfishness.

	C	N
C	1000 , 1000	500 , 900
N	900 , 500	750 , 750

In the case where there are no social preferences, β_i=0, the game has two equilibria: (C, C) and (N, N). The equilibrium (C, C) gives each player the highest possible payoff in the game, but carries some risk: if j decides to play N, then i would get the lowest possible payoff in the game.

There are two reasons why a player would choose N as her strategy. Before delving into them, notice that all players, regardless of their social preferences, prefer to not cooperate when the other player does not cooperate. That is, for all β_i, $u_i(N, N) > u_i(C, N)$.[3] Intuitively, even if i is fully altruistic and has β_i=1, she would still prefer to play N if j plays N because doing so increases her payoff more than it reduces j's payoff (i.e., she increases the sum of both payoffs).

A first reason to choose to not cooperate is because a person has particularly strong hateful preferences. To see this, we need to understand when a person will want to not cooperate even if the other player is going to cooperate. In other words, we analyze the conditions under which $u_i(N, C) > u_i(C, C)$.

$$u_i(N, C) > u_i(C, C)$$
$$900 + \beta_i 500 > 1000 + \beta_i 1000$$
$$\beta_i < -0.2$$

Define the threshold T=-0.2. If i is hateful beyond the threshold, she will prefer to not cooperate regardless of what j will do. That is, if β_i<T, N is a dominant strategy for i. In this particular case, β_i<T means that i is hateful enough to prefer to lose 100 and reduce j's payoff by 500. When β_i<T, we say that i chooses to not cooperate out of hate.

A second reason to not cooperate is because a person fears that the other player may have particularly strong hateful preferences. If i believes that j is hateful beyond the threshold (β_j<T), then i believes that j will not cooperate. And if j will not cooperate, then i will prefer to not cooperate as well, regardless of how altruistic she might be (i.e., $\forall\beta_i$, as shown above). When i believes that β_j<T, we say that i chooses to not cooperate out of fear.

Of course, a person might not be sure if the other player will cooperate or not. Instead, she might believe that there is a certain probability that the other player will not cooperate, given that j is a member of B. Let s_i be i's strategy, and $\tilde{P}_i(s_j$=$N)$ be i's belief about $P(s_j$=$N)$, the probability that j will not cooperate. Then i's expected utility of choosing s_i is:

$$W_i(s_i) = \tilde{P}_i(s_j{=}N) \cdot u_i(s_i, N) \; + \; \tilde{P}_i(s_j{=}C) \cdot u_i(s_i, C)$$

[3] Proof: $u_i(N, N) > u_i(C, N) \Rightarrow 750 + \beta_i 750 > 500 + \beta_i 900 \Rightarrow \beta_i > 5/3$. Because $\beta_i \in [-1, 1]$, it is always the case that β_i<5/3.

Given this, i chooses to not cooperate if $W_i(N) \geq W_i(C)$. Solving for $\tilde{P}_i(s_j{=}N)$ yields the following.

$$\tilde{P}_i(s_j{=}N) \geq \frac{2}{7}\left(\frac{1+5\beta_i}{1+\beta_i}\right) \tag{2}$$

The condition above determines how fearful a person must be in order to not cooperate. Importantly, this depends on i's social preferences. The condition shows that the less altruistic a person is, the less fearful she needs be to want to not cooperate out of fear.

Lastly, a third reason to not cooperate could stem from higher-order beliefs. That is, i's beliefs on j's beliefs, and so on. Consistent with the evidence from the fieldwork and the findings of the experimental literature (Rubinstein, 1989), in the empirical model I will assume that players do not form higher-order beliefs when playing the game. This means that from i's perspective, j will want to not cooperate if, and only if, j is hateful enough ($s_j{=}N \Leftrightarrow \beta_j{<}T$). This has the following important implication:

$$\tilde{P}_i(s_j{=}N) = \tilde{P}_i(\beta_j{<}T)$$

In the Appendix I provide evidence participants in the experiment do not seem to form higher-order beliefs, and that therefore this assumption is the best representation of behavior. In addition, I estimate an upper bound on how big would the role of higher-order beliefs be if this assumption was relaxed.

3.2. Lab experiment design

This section presents the experimental protocol that disentangles the motives of non-cooperation. Following the theory, three key pieces of information are needed from each person in the experiment: (i) the decision to cooperate (s_i); (ii) the social preferences towards the outgroup (β_i); and (iii) the beliefs on the probability that an outgroup member has social preferences below the threshold ($\tilde{P}_i(\beta_j{<}T)$).

To disentangle the primitives that play a role in the equilibrium outcome, I use the following insight when designing the experimental protocol. For each coordination game that participants in the experiment could play, it is possible to design a money allocation decision that mirrors the structure of the coordination game but removes the uncertainty in payoffs, such that beliefs do not enter the empirical problem, only preferences. In this way, the money allocation decision isolates the preferences that play a role in the decision in the coordination game. To clarify the intuition for identification, consider the following example.

	C	N
C	1000 , 1000	500 , 900
N	900 , 500	750 , 750

Coordination Game

↑

preferences

+ beliefs

⟵

	You	Other
Opt 1	1000	1000
Opt 2	900	500

Money Allocation Decision

↑

preferences

In the money allocation decisions, one participant is the decision maker and gets to pick between the payoffs in Option 1 or Option 2, while her match is a receiver. Without imposing any structure on the utility function, we can characterize people into two groups by looking, side by side, at their decisions in the two situations above. If the participant prefers Option 2 to Option 1, she reveals that her preferences are such that she will prefer to not cooperate in the game, even if she thinks that her match will cooperate. We can infer this because the money allocation decision is presenting precisely that scenario. If, instead, the participant prefers Option 1 to Option 2 in the money allocation decision, she reveals that she would prefer to cooperate in the game if she thinks her match will cooperate. Therefore, if a participant prefers Option 1 to Option 2, but in the coordination game decides to not cooperate, it must be because she believes that her match will not cooperate.[4] As I will explain below, in the experiment participants will face multiple money allocation decisions to capture with more precision their level of social preferences. I will also use the money allocation decisions to elicit beliefs about others' social preferences.

Experiment set up

With the insight from above in mind, the lab experiment proceeds in the following way. First, participants are told there are two groups, the Blue Group and the Green Group, and that they will be randomly matched and play with one person from each group (although they will not know whom from each group, just their match's group membership). Then, participants are shown the list of names of the people that belong to one of the groups (the group they face first is picked at random). Each group consists of ten people. Crucially, the names in the Blue Group are all Christian names, while the names in the Green Group are all Muslim names. In Nigeria, names are a clear signal of religious affiliation, so participants are able to easily identify that the common characteristic of the members of each group was religion. After playing some games with the group that was first revealed, participants are then shown the list of names of the second group and proceed to play with that group. By the end of the experiment, participants have gone through the same activities with their match from each group.

This design has two advantages. First, using names to signal religions allows me to not mention religion explicitly, which helps reduce experimenter demand bias[5]. Second, by not knowing exactly

[4] If no structure is imposed on the utility function, an underlying assumption in this analysis is that if a person believes the other player will not cooperate, they will prefer to not cooperate too.

[5] In the results section I also check how robust is the analysis to controlling for social desirability bias.

who their match is within the group, participants are forced to think about the average behavior of the members of the group, of which the only discernible shared characteristic is religion. In this way, one can control for any change in the participant's behavior due to interacting with a male or female name, or names that are probably from older or younger cohorts.

Importantly, there was no deception in this experiment—the names in the Blue and Green groups belonged to real people, and the payoffs of the games were real, for both participants and receivers.

Eliciting social preferences

After participants are matched with an unknown person from the first group shown to them, they start the activities of the experiment with a series of money allocation decisions. There were 20 money allocation decisions a participant could potentially face. I designed an algorithm to elicit social preferences with the lowest number of questions possible. In the end, participants face 7 or 8 money allocation decisions with each match. In half of the 20 money allocation decisions a participant could face, Option 2 represents the hateful option of reducing the match's payoff in 500 naira, for a price. In the other half, Option 2 represents the altruistic option increasing the match's payoff in 500 naira, for a price. Within each half, each possible decision varies the amount of money a person has to give up to pick Option 2 (i.e., each decision presents a different price for Option 2). In the end, this series of decisions allows to elicit the participant's willingness to pay to either increase or decrease in a fixed amount the payoff of their match. Full details on this design can be found in the Appendix.

The choices in the money allocation decisions sort participants into one of 21 types. Assuming the utility functional form of equation (1), I can assign to each type a calibrated social preference parameter using the following result: a money allocation decision where a participant picks Option 2 reveals that $\beta_i \leq (x_{i2} - x_{i1})/(x_{j1} - x_{j2})$—where x_{i1} is the payoff for participant i if she picks Option 1.[6] Using this calibration method, I can place each participant's social preference parameter in one of the following preference intervals: $\hat{\beta}_i \in \{(-1, -0.9), ..., (0.9, 1)\}$. This calibration approach has the advantage of being simple and transparent. However, a downside is that it ignores sampling variability, so standard errors of the estimated individual-level parameters can not be calculated. Section 4 presents an alternative approach that addresses this problem at the cost of being less direct.

Eliciting beliefs about others' social preferences

After the money allocation decisions phase, the next module elicits beliefs on the probability of j not cooperating. Recall from 3.1 that $\tilde{P}_i(s_j{=}N) = \tilde{P}_i(\beta_j{<}T)$. Therefore, it is enough to elicit i's belief on the probability of j being hateful beyond the threshold to know i's belief on the probability of j not cooperating.

[6] This expression results from the following process. Let the utility of picking Option 1 be $u_i(Opt1) = x_{i1} + \beta_i x_{j1}$. The utility of picking Option 2 is analogous. Then, choosing Option 2 means that $u_i(Opt2) \geq u_i(Opt1)$. That is, $x_{i2} + \beta_i x_{j2} \geq x_{i1} + \beta_i x_{j1}$. Solving for β_i we get $\beta_i \leq (x_{i2} - x_{i1})/(x_{j1} - x_{j2})$.

With this in mind, I ask participants to guess the choices that other participants made in the money allocation decisions. If they guess correctly, participants get extra payment. Notice, first, that beliefs on $P(\beta_j<T)$ are determined by the beliefs on the distribution of social preferences of the group j belongs to. Put differently, $\tilde{P}_i(\beta_j<T)$ is determined by i's beliefs on the distribution of $\beta_j|j\in G$ (with *G=Green,Blue*). I elicit beliefs on the mean of this distribution and the mass of the tail at key points that are directly connected to coordination games that participants will play afterward.

First, participants go through the series of money allocation decisions again, trying to guess what their match from the Green/Blue group picked. With this, I elicit the beliefs on the mean, $\tilde{E}_i[\beta_j]$. Then, I ask participants to guess how many people (out of the 10) from the Green/Blue group picked Option 2 in two particular money allocation decisions. Each of these scenarios represents a threshold where someone would reveal to be hateful enough to want to not cooperate out of hate in an upcoming coordination game. Ultimately, this question elicits the beliefs on the percentage of Muslims/Christians that are hateful enough to want to not cooperate in an upcoming coordination game—that is, it elicits $\tilde{P}_i(\beta_j<T)$ given that $j\in G$, which in the end elicits $\tilde{P}_i(s_j=N)$.

Measuring cooperation

Lastly, participants play coordination games with their anonymous matches from the Green and Blue groups. With each match, they play two coordination games, where each game has a slight variation in payoffs that changes the threshold of how hateful a person needs to be to want to not cooperate out of hate. After these, the activities of the lab conclude. Further details on the protocol can be found in the Appendix, where the reader can also find details of the coordination games and money allocation decisions, and screenshots of how these were presented to participants.

4. Empirical Model and Estimation

While simple and transparent, the approach presented in Section 3.2 to estimate the social preferences of participants has some drawbacks. First, individual-level parameters are being calibrated using 7 or 8 decisions per person. This procedure ignores sampling variability and, therefore, does not allow estimating standard errors of the estimated preference parameters. Additionally, the first approach does not allow to test alternative models that introduce other parameters that could explain the decisions in the game, like loss aversion or psychological costs. This section presents an estimation procedure that overcomes these problems and still allows to recover parameters at the individual level to determine the extent to which non-cooperation is driven by hate vs. fear.

In what follows I introduce an empirical model with random coefficients to recover $\beta_i\ \forall i$. In short, this procedure uses everyone's full set of decisions to estimate the distribution from where β_i's are drawn, and then uses an individual's decisions to determine where in the estimated distribution the individual's β_i is likely to be. To simplify the explanation in this section, I will focus on the case where a partici-

pant $i \in A$ is matched with a participant j who belongs to the outgroup, $j \in B$. But notice that the same parameters can be calculated for the case where j belongs to the ingroup, $j \in A$.

In the experiment, participants are matched with an unspecified $j \in B$. They make M_i money allocation decisions, where M_i can be 7 or 8 depending on the participant's decisions. Their beliefs on the money allocation decisions j made are elicited. And they play G coordination games, with $G=2$.

In each money allocation decision m, participant i makes decision d_{im} between two options with sure payoffs for herself and her match j. Participant i's utility of picking $d_{im} \in \{Opt1, Opt2\}$ is her base utility function (as defined in 3.1), plus an error, ε_{id}, that has an extreme value distribution with mean zero. This error can be thought of as the result of limited attention in the experiment. The utility function is:

$$u(d_{im}) = x_{im}^{d_{im}} + \beta_i \cdot x_{jm}^{d_{im}} + \varepsilon_{id}$$

The data consists of is d_{im} and the payoffs for i and j in each option of each money allocation decision. The unknown parameter is β_i. Because ε_{id} is distributed extreme value, the probability of participant i's sequence of choices $d_i = \langle d_{i1}, ..., d_{iM_i} \rangle$ is:

$$\Lambda_{im} = \frac{\exp\Big(u(Opt2) - u(Opt1)\Big)}{1 + \exp\Big(u(Opt2) - u(Opt1)\Big)}$$

$$P(d_i|\beta_i) = \prod_{m=1}^{M_i} \Lambda_{im}^{\mathbb{1}(d_{im}=Opt2)} \left(1 - \Lambda_{im}\right)^{\mathbb{1}(d_{im}=Opt1)}$$

Participants also play G coordination games. In each game g, participant i picks strategy $s_{ig} \in \{C, N\}$. Participant i has risk-neutral preferences and her expected utility function includes an error, ε_{is}, that has an extreme value distribution with mean zero, and that is independent from ε_{id}. The expected utility function is:

$$W(s_i) = \tilde{P}_i(s_j{=}C) \cdot u(s_i, s_j{=}C) \; + \; \tilde{P}_i(s_j{=}N) \big[u(s_i, s_j{=}N) - \psi_i \cdot \mathbb{1}(s_i{=}C)\big] + \varepsilon_{is}$$

$$u(s_i, s_j) = x_i^{s_i, s_j} + \beta_i x_j^{s_i, s_j}$$

Where $\tilde{P}_i(s_j{=}s)$ is i's subjective beliefs on $P(s_j{=}s)$, given that $j \in B$. Recall that $\tilde{P}_i(s_j{=}N) = \tilde{P}_i(\beta_j{<}T)$, and $\tilde{P}_i(\beta_j{<}T)$ is elicited directly in the experiment (see section 3.2 for details).

ψ_i is a parameter of loss aversion, where the reference point is the payoff from *(Cooperate, Cooperate)*. In addition, this parameter can also capture a distaste for getting what is usually described as the "sucker's payoff" (the payoff i gets when she cooperates and j does not). Including this loss aversion parameter relaxes the functional form assumption and gives more flexibility to the model to better fit the data.[7]

[7] An alternative approach to modeling this situation would be to have an expected utility function with risk aversion. I discard this approach because at this level of prices individuals should not exhibit risk aversion. Indeed, this is what I find in

In the Appendix I test for alternative functional forms and find this to be the best one (according to the Likelihood Ratio Test). Importantly, notice that ψ_i is multiplied by $\tilde{P}_i(s_j{=}N)$, so not cooperating because of ψ_i also means not cooperating out of fear that j will not cooperate. ψ_i and β_i are independent.

The data consists of s_{ig}, $\tilde{P}_i(s_{jg}{=}N)$, and the payoffs for i and j in all four scenarios of each game. The unknown parameters are β_i and ψ_i. Because ε_{is} is distributed extreme value, the probability of participant i's sequence of choices $s_i = \langle s_{i1}, s_{i2} \rangle$ is:

$$\Lambda_{ig} = \frac{\exp\Big(W(N) - W(C)\Big)}{1 + \exp\Big(W(N) - W(C)\Big)}$$

$$P(s_i|\beta_i, \psi_i) = \prod_{g=1}^{G} \Lambda_{ig}^{\mathbb{1}(s_{ig}=N)} \left(1 - \Lambda_{ig}\right)^{\mathbb{1}(s_{ig}=C)}$$

Combining both probabilities, I can define the probability of i's sequence of choices in all the lab games, $y_i = \langle d_{i1}, ..., d_{iM_i}, s_{i1}, s_{i2} \rangle$:

$$P(y_i|\beta_i, \psi_i) = P(d_i|\beta_i) \times P(s_i|\beta_i, \psi_i)$$

Let $\theta_i \equiv (\beta_i, \psi_i)$, a vector of our parameters of interest. I assume that $\theta_i \sim \mathcal{N}(\mu, \Sigma)$ and has a probability density function $f(\cdot)$. So the probability of i's sequence of choices y_i is:

$$P(y_i|\mu, \Sigma) = \int P(y_i|\theta_i) \cdot f(\theta_i|\mu, \Sigma) \ d\theta$$

A mixed logit likelihood function represents the probability of observing all the decisions of all individuals:

$$L = \prod_{i=1}^{N} P(y_i|\mu, \Sigma)$$

Because the integrals in the likelihood function are hard to calculate, they are approximated through numerical simulations. The parameters μ and Σ are estimated through simulated maximum likelihood, following Train (2009).

After estimating μ and Σ, I can use them to subsequently estimate $\theta_i \ \forall i$. Using Bayes' rule, I can derive a distribution of θ_i conditional on i's sequence of choices y_i:

$$g(\theta_i|y_i, \mu, \Sigma) = \frac{P(y_i|\theta_i) \cdot f(\theta_i|\mu, \Sigma)}{P(y_i|\mu, \Sigma)}$$

Using $g(\cdot)$, I can calculate the mean of the distribution conditional on the choice sequence y_i, and

the field. Using a canonical survey module to measure risk aversion, I find that over 90% of individuals are risk-neutral. In addition, the behavioral literature suggests that at low prices, behavior is better explained by loss aversion than risk aversion (Rabin, 2000; DellaVigna, 2018).

use it as an estimator of θ_i:

$$\bar{\theta}_i = \int \theta_i \cdot g(\theta_i | y_i, \mu, \Sigma)$$

This integral is approximated through simulations, following Train (2009).

It is worth noticing that this estimation procedure manages to use all the information in one single stage while keeping the essence of the identification strategy of the experimental design, which is to estimate social preferences separately from the coordination games. In this estimation, 80% of the observations used to estimate μ_β come from the money allocation decisions. Intuitively, what the estimation will tend to do is to pick a μ_β to fit the money allocation decisions, and pick a μ_ψ to fit the coordination game decisions that remain unexplained, given the individual beliefs imputed, $\tilde{P}_i(s_j{=}N)$.

5. Policy analysis: RCT of a radio drama

After understanding the motives behind social failure, the next step is to study whether policy interventions used in this setting are increasing cooperation and why. The policy I analyze is the production of media content to promote intergroup cooperation. In particular, I focus on radio dramas, a policy that has been widely popular in Sub-Saharan Africa. In Nigeria, NGOs are constantly creating new radio dramas to promote messages on different topics. For example, the main production company in the Jos region creates around 4 radio dramas per year, and recently created shows on topics such as women empowerment and Covid-19. Moreover, radio dramas have been used to promote messages on conflict-related issues. For instance, radio dramas have tackled topics such as how fake news fuels conflict and the reintegration of former Boko Haram members into society.

In Nigeria, policymakers view radio dramas as a highly valuable strategy for addressing conflict, citing three primary reasons. First, fictional stories make it easier to discuss sensitive topics (Slater and Rouner, 2002). Delving into historical and contemporary conflict tends to evoke strong emotions in the listeners, which can make them less receptive to the intended message. A fictional story overcomes this challenge. Second, dramatized narratives help to increase attention and retention of the intended message (Kromka and Goodboy, 2019). In an environment saturated with numerous NGOs constantly employing different sensitization campaigns to promote cooperation, novel initiatives struggle to capture people's attention. Instead, radio dramas stand out due to their engaging nature, and using narratives has been shown to increase message retention. Third, when compared to alternative policies on conflict, radio dramas can be easily implemented in a wide range of contexts. Lately, intergroup contact interventions have received considerable attention. But these policies can only be carried out in very particular contexts, where the two communities in conflict live together and tensions are not such that intervention could lead to violence. Instead, radio dramas can be implemented in places where only one of the two groups in conflict live. In addition, they are relatively low cost and require minor logistics.

These three reasons help explain why radio dramas have become a popular policy in Africa and

make them an interesting policy to study. In addition, research in social psychology on whether radio dramas can improve relationships between groups after conflict has found mixed results (Paluck, 2007), indicating the need for further investigation. My goal is to evaluate this policy using the experimental protocol in order to understand its effects on cooperation, hate and fear.

5.1. The radio drama

I partnered with a radio drama production company from Nigeria to create a *new* radio drama. This company has been hired to create the radio shows of some of the most important NGOs in Nigeria, like Search for Common Ground and UN Women. Creating a new radio show has important advantages. First, it ensures that the participants of the experiment have not previously heard the treatment radio show. Using an existing radio show would pose a problem because these are widely broadcasted, which means that the subjects of the experiment could have already been treated. Alternatively, one could use a radio show that was broadcasted in an area that does not cover Jos. However, importing a radio drama would not have the same effect would not be as effective since the messages of this type of radio dramas are tailored to the specific situation of the place in which they are broadcasted. Moreover, creating a new radio show allowed me to have a story that directly addressed the motives I explore in this paper—that is, a story that spoke about hate and fear between communities in conflict.

A possible concern about creating a new radio show is that one might not be evaluating the exact same policy implemented by policymakers. On this, it is important to note that even though the NGOs pay for the shows, the creative process relies on the production company. To emulate the policy creation process as closely as possible, I follow the exact same steps Nigerian NGOs take to create their radio shows. These steps are straightforward. (i) The NGO hires the production company to create a new radio show. (ii) The NGO provides one page of pointers stating the main message they want the show to convey. (iii) The production company gets back to the NGO with an outline of the story and how it conveys the message, and the NGO approves or makes comments. (iv) The production company writes the scripts for the episodes and sends them to the NGO for approval. (v) The production company records the show and delivers the final product to the NGO.

The objective of the treatment was to reduce both hate and fear, as the treatment was to be designed before knowing which driver of conflict was best to focus on (emulating real-life policy design). With this in mind, the pointers I gave to the production company were the following. I wanted a story that promoted interfaith peace and cooperation. The story should showcase two communities in conflict where hate and unwarranted fears lead both communities to miss out on mutually beneficial interactions. The resolution of the story should convey a message on how reevaluating fear and letting go of hate can lead to both communities being better off.

The radio drama that was created is called *The Convergence*. It consisted of 24 episodes lasting between 10 to 15 minutes, was available in both English and Hausa and participants could listen to

whichever they preferred. The plot unfolds as follows: A corrupt politician offers contracts to a businessman in exchange for ensuring his victory in the election. To achieve this, the businessman assembles a team to disseminate fake news on social media, fueling tribal conflicts and creating fear to discourage people from voting. This leads to rising tensions and unfair judgments about the outgroup, which result in important important losses for the communities, including the unjust firing of an outstanding schoolteacher due to her tribal affiliation and the rejection of a beneficial NGO program solely because its leader came from a different tribe. Moreover, one key character harbors deep resentments toward the other tribe because of past family tragedies caused by the ongoing conflict. As the story reaches its resolution, the communities uncover the politician's manipulation scheme and vote him out of office; the businessman flees the country, while his collaborators face imprisonment; and the character with hate has a healing process, enabling her to form meaningful friendships with members of the other tribe, fostering reconciliation and unity.

5.2. RCT design

The RCT for the radio show was conducted in the following way. Initially, subjects were recruited to participate in two lab-in-the-field experiments, two months apart. To reduce demand effects, only at the end of the baseline lab experiment enumerators asked participants if they were interested in participating in "a different project the surveying company was carrying out." They told participants this project consisted of listening to a new radio drama a production company was releasing and providing feedback on it. Participants were also informed that they could listen to it sometime before the second lab experiment they had agreed to take. Those who were interested were invited to sign up immediately and were told that they would receive more details in the following days through WhatsApp (using the contact number collected for the second lab experiment.)

Individuals were randomly assigned to either the treatment or control groups. The show was not broadcasted, but instead episodes were sent to participants through WhatsApp four times a week (on Mondays, Wednesdays, Fridays and Saturdays) over a six-week period. To promote and monitor engagement, every Saturday participants received a quiz on the content of that week's episodes. Answering the quiz correctly put people in a weekly lottery for two prizes of 2,000 naira, and gave them one entry to the two grand prizes of 50,000 naira, which were awarded at the end of the sixth week. The quizzes also asked for participants' opinions on the radio show. The control group was sent a placebo radio show with a message on health. They also received weekly quizzes with the same scheme of prizes.

A week after the radio show ended, the endline lab-in-the-field experiment started. Participants in the treatment and control groups went through the lab experiment again, which allowed me to remeasure their preferences, beliefs and cooperation to estimate the effects of the radio drama on each margin.

The treatment and control groups were balanced on baseline levels of cooperation, social prefer-

ences, beliefs, religion, sex, age and other characteristics. A balance table can be found in the Appendix. The attrition rate at endline was 5%.

5.3. Data collection

The fieldwork of this project took place between December 2022 and February 2023. Data collection for each lab-in-the-field (baseline and endline) lasted for two weeks. The treatment took place for six weeks, in between surveys. At baseline, the team in the field surveyed 997 people from 41 Jos communities (out of 44 communities). The sample was 50% Muslim and 50% Christian, 47% female and 53% male, with ages between 18 and 60 and a mean of 33. Participants were required to have access to a phone with WhatsApp and be available to participate in a second lab experiment two months later.

The recruiting process was the following. Every morning a pair of enumerators of the same religion visited a community of their religion. When in the community, enumerators picked a random starting point (like a school or water source) and started walking in opposite directions. To select a house to survey, they followed a 3/4 pattern, knocking on the 3rd house away from the starting point, then the 4th house from there, then the 3rd house from there, and so on. If someone answered the door, enumerators would briefly explain the survey and ask if someone in the household was interested in participating. If someone accepted, the lab-in-the-field experiment was carried out immediately at the person's home. Enumerators were instructed to maintain a balanced sample in age and gender. On average, the survey took around 45 minutes to complete.

At the end of the survey, enumerators asked participants if they wanted to participate in the radio show project. Because this implied no extra effort, everyone agreed to be contacted for this. Some days after the baseline was completed, we created two WhatsApp groups, one for the treatment group and another one for the control group. In them, we welcomed everyone to the radio show project and explained the logistics of it. Through the WhatsApp groups we sent the episodes of the radio drama and the link to the weekly quizzes, and announced the winners of the prizes. Only administrators could send messages in these WhatsApp groups.

After the radio show ended, enumerators visited the communities again. Using the registered phones, enumerators contacted the participants and scheduled appointments to carry out the endline lab experiment. 947 participants from baseline participated in the endline lab experiment—an attrition rate of around 5%.

For each survey, participants received a compensation between 700 and 1,700 naira, depending on the results of the different lab games. In Jos, 1,000 naira is approximately the payment for four hours of work. These payments were made in cash immediately after the survey ended. Because only the payoff some questions, picked at random, were implemented, the final payment did not give any information about the participant's (or her match's) decisions. The payments of the quiz lotteries were made directly to winners' accounts via phone transfer as soon as the winners were announced.

6. Results

6.1. Descriptive evidence

The first result is that, in a game where cooperating is both an equilibrium and the maximum payoff (which represents half a day of salary), people fail to cooperate in 31% of the interactions between groups—compared to only 6% of interactions within groups. This results in a 9.6% loss of total wealth in intergroup interactions.

Figures 2A and 2B displays three main diagnostic facts drawn from the baseline lab-in-the-field experiment. The social preferences parameters shown in these figures were estimated following the approach described in Section 3.2.

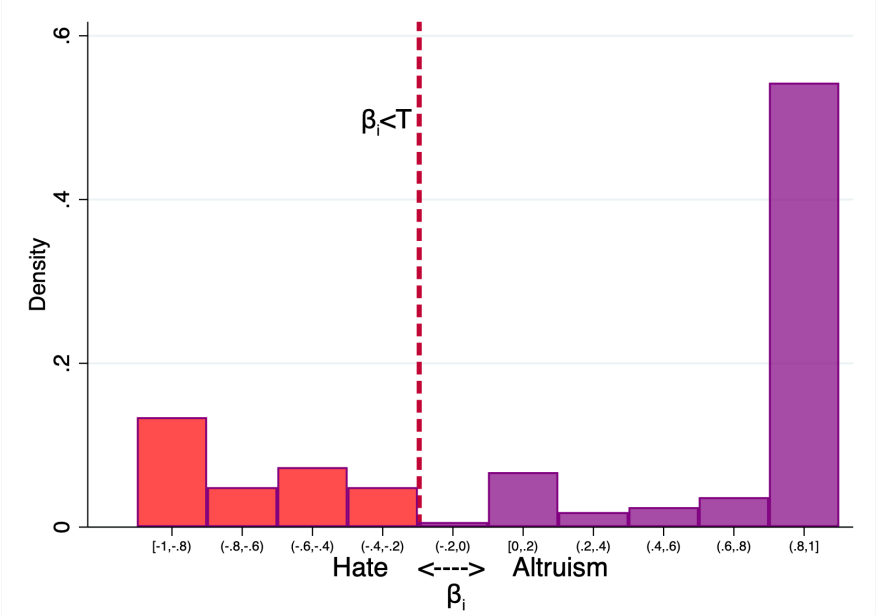
Figure 2A presents a histogram of the social preferences of the participants who decided to not cooperate in the main coordination game of the experiment. The x-axis shows social preferences ranging from fully hateful, $\beta_i{=}{-}1$, to fully altruistic, $\beta_i{=}1$. The dashed red line represents the threshold point where a person becomes hateful enough to want to not cooperate out of hate. Figure 2A provides an initial classification of motives for non-cooperation. The 30% of non-cooperators who fall to the left of the dashed line chose to not cooperate out of hate. On the other hand, the 70% that fall to the right did not have a hateful motive to not cooperate, but did so out of fear. It is also worth noting that rightmost bar indicates that 52% of non-cooperators are, in fact, highly altruistic towards the outgroup.

Figure 2B displays a histogram of the beliefs about the outgroup, for cooperators and non-cooperators. Specifically, it shows what people believe is the likelihood that an outgroup member wants to not cooperate out of hate. That is, beliefs about the percentage of outgroup members that have a level of hateful beyond the threshold, $P(\beta_j{<}T)$. The dashed blue line shows the actual probability of this event happening, which is 6%. On average, cooperators believe that 14% of the outgroup will not cooperate out of hate, while non-cooperators believe 59% will do so. This means that non-cooperators exaggerate the number of hateful people in the outgroup by around 10 times. This fact is evidence of how unfounded fears play a central role in cooperation failure.

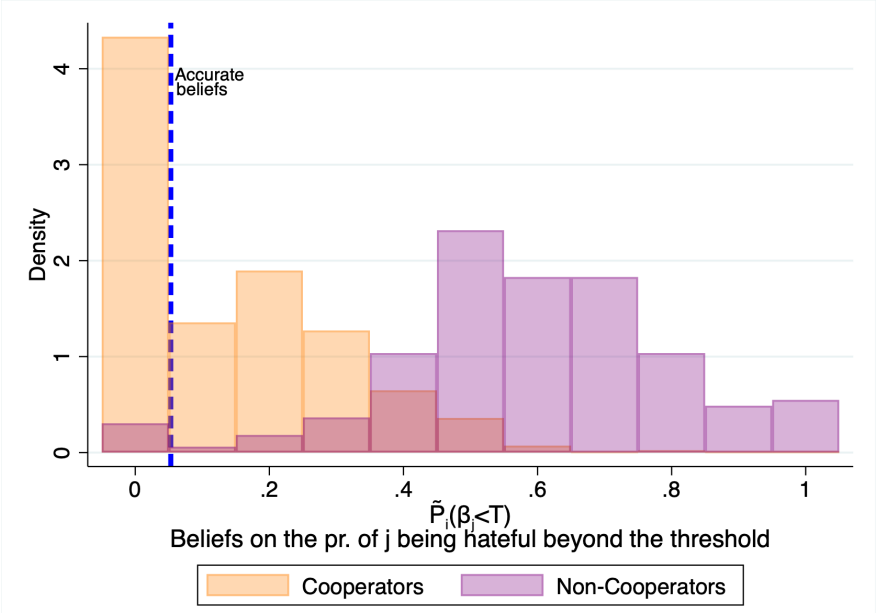
Figure 3 shows the relationship between preferences, beliefs and cooperation for the entire sample. The x-axis shows participants' social preferences towards the outgroup, and the y-axis shows participants' beliefs about the outgroup's social preferences towards the ingroup (more specifically, beliefs on the percentage of outgroup members that are not hateful beyond the threshold). Each participant is represented by a dot. The dot is green if the participant cooperated and red if they did not. The dots are translucent and can overlap. This creates different grades of opacity, with darker dots indicating a higher density of people at that preference-belief level. In addition, the greener a dot is, the more cooperation there is at that preference-belief level, and the redder it is, the more non-cooperation there is. (Brown dots are the result of translucent green and red dots overlapping.) The black line represents the fitted values. The vertical dashed line represent the threshold, $T{=}{-}0.2$, at which an individual becomes hateful enough to want to not cooperate out of hate. To the right of this line, when an individual does

Figure 2

A. Social Preferences (for Outgroup) of Non-Cooperators



B. Beliefs about the Outgroup, by Game Strategy



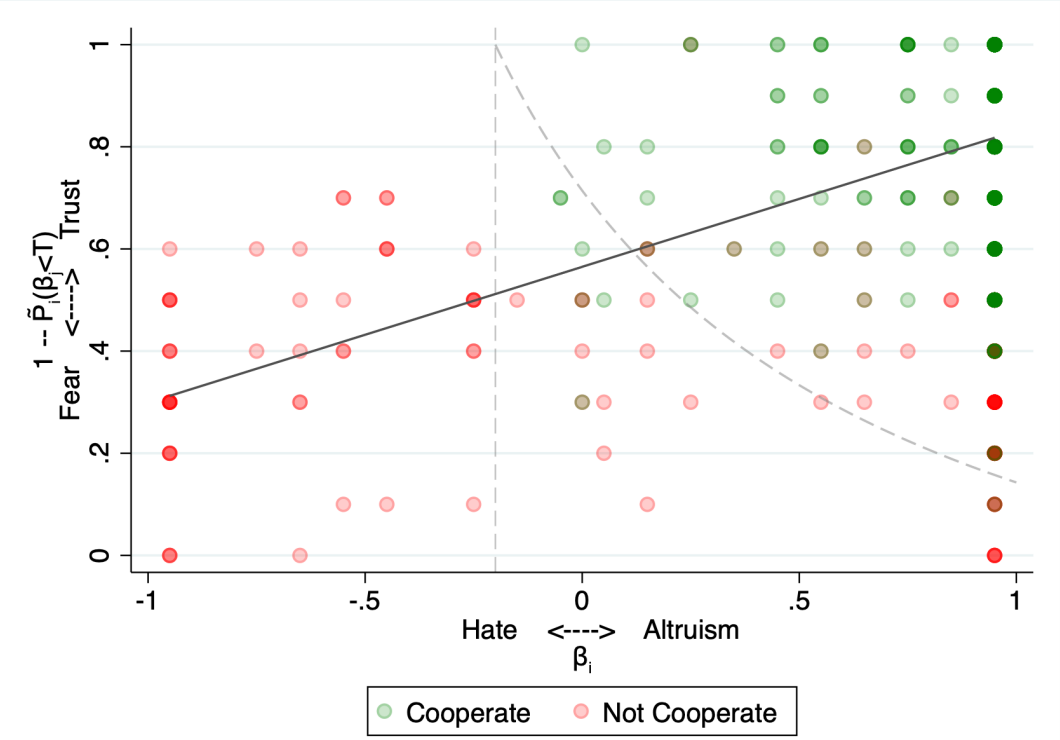
Notes: Figure 2A presents the social preferences (β_i) of non-cooperators, estimated following the approach described in Section 3.2. The red line represents the threshold point where a person becomes hateful enough to want to not cooperate out of hate. Figure 2B displays the beliefs about probability that an outgroup member has a level of hate beyond the threshold, $P(\beta_j<T)$, and therefore want to not cooperate out of hate. The blue line represents the actual probability of this event happening.

not have a hateful motive to not cooperate, the curved dashed line represent the threshold below which an individual is fearful enough to want to not cooperate out of fear (which depends on β_i, as described by equation (2)).

Figure 3 has a few facts worth noticing. First, no one to the left of the hate threshold ($\beta_i<T$) cooperates. Second, when $\beta_i>T$, as fear increases, the rate of non-cooperation increases. Third, on average

there is a positive correlation between preferences and beliefs. Fourth, this correlation is stronger on the hateful side than on the altruistic side: hateful individuals tend to be also very fearful, whereas altruistic individuals show a wider range of beliefs, from fully trusting to fully fearful.

Figure 3: Preferences, Beliefs and Cooperation



Notes: This figure shows the relationship between preferences, beliefs and cooperation for the entire sample. The x-axis shows participants' social preferences towards the outgroup, and the y-axis shows participants' beliefs about the outgroup's social preferences towards the ingroup (more specifically, beliefs on the proportion of outgroup members that are not hateful beyond the threshold). Each participant is represented by a dot. The dot is green if the participant cooperated and red if they did not. The dots overlap and are translucent. This creates varying opacities, with darker dots indicating a higher density of people at that preference-belief level. In addition, the greener a dot is, the more cooperation there is at that preference-belief level, and the redder it is, the more non-cooperation there is. The black line represents the fitted values. Social preference showcased here were estimated following the approach described in Section 3.2.

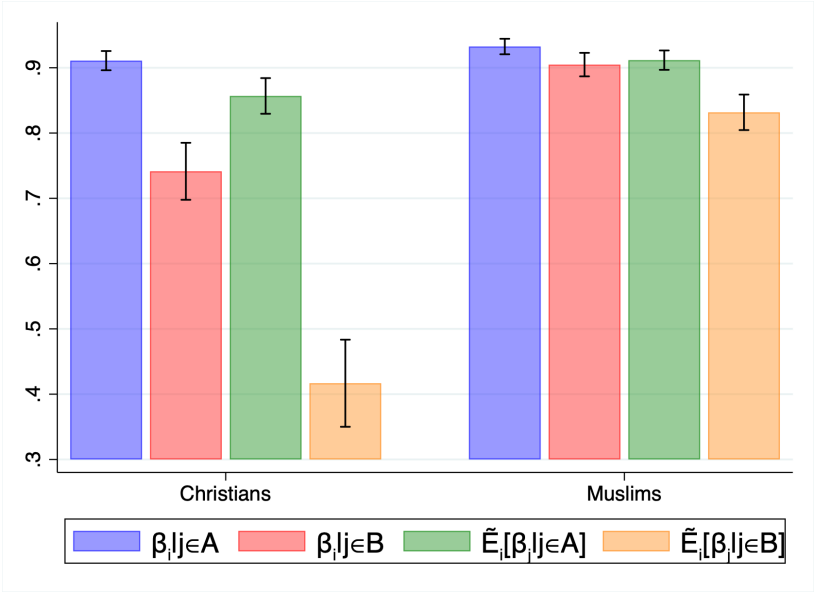
Heterogeneity by religion

Do the two religious groups under study act in a similar way? Or if there is heterogeneity in behavior, in which direction does it go? In terms of cooperation, the difference is stark. Out of all the people who decided to not cooperate with the outgroup, 84% were Christians, while only 16% were Muslims. Figure 4 reports heterogeneity in social preferences and beliefs.

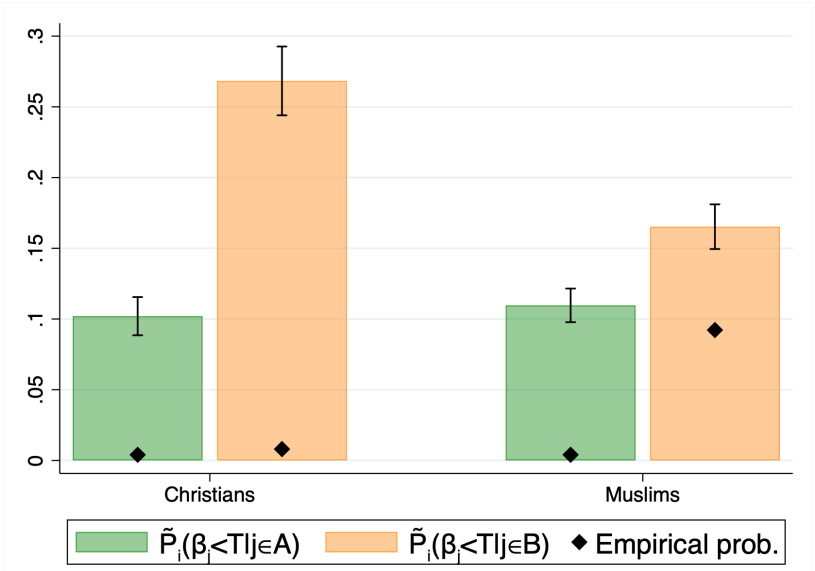
Figure 4A reports the following averages, from left to right: social preferences for the ingroup ($\beta_i|j\in A$), social preferences for the outgroup ($\beta_i|j\in B$), beliefs on the mean social preferences ingroup members have for the ingroup ($\tilde{E}_i[\beta_j|j\in A]$), beliefs on the mean social preferences outgroup members have for the ingroup ($\tilde{E}_i[\beta_j|j\in B]$)—this for Christians and Muslims. When comparing the blue and the red bars, the figure shows that both Christians and Muslims have more positive social preferences with the ingroup than the outgroup. However, the gap between average social preferences for the ingroup

Figure 4: Heterogeneity by Religion

A. Preferences and Beliefs about the Mean



B. Beliefs about the Tail



Notes: In both figures, $j \in A$ means that j belongs to the ingroup, and $j \in B$ means that he belongs to the outgroup. Figure 4A reports the following averages, from left to right: social preferences for the ingroup ($\beta_i | j \in A$), social preferences for the outgroup ($\beta_i | j \in B$), beliefs on the mean social preferences ingroup members have for the ingroup ($\tilde{\mathrm{E}}_i[\beta_j | j \in A]$), beliefs on the mean social preferences outgroup members have for the ingroup ($\tilde{\mathrm{E}}_i[\beta_j | j \in B]$). Figure 4B presents beliefs about the tail of the distribution of social preferences, for Christians and Muslims. $\tilde{P}_i(\beta_j < T | j \in A)$ is the beliefs on the percentage of ingroup members that are hateful beyond the threshold. $\tilde{P}_i(\beta_j < T | j \in B)$ is the beliefs on the percentage of outgroup members that are hateful beyond the threshold. The diamonds represent the actual percentage of people that are hateful beyond the threshold for each case. In both figures, the black lines represent standard errors.

vs. the outgroup is much greater in Christians than in Muslims. In addition, comparing both red bars shows how Christians have worse social preferences toward Muslims than Muslims toward Christians. In terms of beliefs about the ingroup, comparing the green and the blue bars shows that both groups have close to accurate (although a little pessimistic) beliefs on how ingroup members treat other ingroup

members. Finally, regarding the beliefs about the outgroup, comparing the yellow bar of one group with the red bar of the other group reveals an important difference between religious groups: Christians have very inaccurate and pessimistic beliefs about how Muslims treat them. Instead, if anything, Muslims have somewhat optimistic beliefs on how Christians treat them.

Figure 4B presents more evidence of the heterogeneity in misperceptions between groups by looking at beliefs about the tail of the distribution of social preferences—that is, beliefs about the percentage of people that are hateful beyond the threshold, $\tilde{P}_i(\beta_j<T)$. Blue bars report beliefs about the ingroup and red bars beliefs about the outgroup. The first thing to note is that both groups exaggerate the percentage of hateful people there is in both the ingroup and the outgroup. In terms of perceptions about the ingroup, the bias is very similar between the groups—both group believe that around 10% of the ingroup is hateful towards other ingroup members, while in reality it is only 1%. However, regarding perceptions about the outgroup, the bias differs considerably. Christians believe that 26% of Muslims are hateful beyond the threshold, while Muslims believe 16% of Christians are. In addition, because there are fewer hateful Muslims than hateful Christians, Christians exaggerate the amount of hateful people in the outgroup by 24 percentage points, while Muslims exaggerate this amount by 7 percentage points.

In sum, in this context, Christians are less cooperative than Muslims, have worse social preferences towards Muslims and have more biased beliefs about Muslims than the other way around. Importantly, in the pre-analysis plan I pre-registered that heterogeneity in these three outcomes would go in this direction, based on the focus groups done in the exploratory phase of this project. The main reason why this is the case is probably the salience of Boko Haram, the major armed group in the country, which distorts the beliefs of only one group and generates negative feelings towards only one group. However, it is important to recall that religious violence in this are has come from both sides. Another feature of this context that probably helps drive the heterogeneity in behavior is the location of Jos. Although the city has a very similar number of Christians and Muslims, most of the nearby cities outside of Plateau State are part of the Muslim north. This creates the feeling in some that the Christians in Jos are surrounded by Muslim communities and that therefore their presence in the area is threatened. Importantly, Christians and Muslims have similar levels of income in this setting, so this should not be driving the results.

6.2. Estimated model and conterfactual analysis

Table 1 reports the results of the simulated maximum likelihood estimation of the random coefficients model presented in Section 4. The parameters of interest are β_i, the social preferences for the outgroup, and ψ_i, the loss aversion, both for all i. I estimate the mean and variance of the distribution where these parameters are drawn from.

The first thing to note is that all parameters are precisely estimated—all four parameters in Table 1 are significant at the 1% level. The estimated μ_β shows that, on average, people are highly altruistic

Table 1: Random Coefficients Estimation

	Coefficient	Stand. Err.	
μ_β	0.922	0.072	***
σ_β	0.420	0.059	***
μ_ψ	532.7	108.6	***
σ_ψ	469.2	163.7	***
Observations			9,006
Clusters			997

Notes: This table reports the results of the simulated maximum likelihood estimation of the random coefficients model presented in Section 4. Each observation is one decision of one participant in either a money allocation decision or a coordination game. μ_β and σ_β are the mean and variance of the distribution of the parameters of social preferences. μ_ψ and σ_ψ are the mean and variance of the distribution of the parameters of loss aversion. Standard errors are clustered at the individual level.

towards the outgroup. In addition, σ_β indicates the level of dispersion of social preferences around the mean. The estimated μ_ψ shows that the loss aversion (or/and the psychological cost of getting the "sucker payoff") matters when playing the game and is of considerable size, being half the amount of the maximum payoff in the game. Nevertheless, the size of σ_ψ highlights how this penalty varies considerably in the population.

The fact that $\psi_i > 0$ warrants some discussion. The reason why this is the case is that half of the people who did not cooperate are fully altruistic, as shown in Figure 2. If ψ_i=0, a fully altruistic person would want to not cooperate only if she believes that $P(s_j{=}N){>}0.86$. However, fully altruistic non-cooperators believe, on average, that $P(s_j{=}N){=}0.6$. This implies that the potential costs of cooperating go beyond the monetary one. Alternatively, this gap could imply that higher-order beliefs are formed and play a role in the cooperation decision. In the Appendix I show evidence that this is unlikely to be the case. In addition, I show that in the extreme case where participants had no risk aversion, loss aversion or psychological costs, higher-order beliefs would account for *at most* 27% of the fear driving non-cooperation.

Using the estimated distributions, I estimate individual-level parameters to assign a β_i and ψ_i to each participant, following the procedure explained in Section 4. I use these parameters and the elicited beliefs in the analysis that follows. The first important thing to note is that the model performs well in terms of sample fit: using the estimated individual-level parameters and the elicited beliefs, I correctly predict 94% of the decisions in the coordination game. This is indication that the core drivers of non-cooperation are captured by my model and measurements.

To determine to what extent non-cooperation is driven by hate vs. fear, I shut down the fear motive in participants' expected utility and observe how many non-cooperators would still prefer to not cooperate in this scenario. By doing this I determine what percentage of people do not cooperate purely out of hate and what percentage require fear to decide to not cooperate. More specifically, to shut down the fear channel I set to zero participants' beliefs on the probability that j will not cooperate—that is, I set

$\tilde{P}_i(s_j{=}N){=}0$ for all i. Doing so reduces the expected utility to $W(s_i){=}u(s_i, C)$. Because I have estimated β_i for all i, I can calculate $u(s_i, C)$ and determine for each i if $u(N, C){>}u(C, C)$. As shown before, $u(N, C){>}u(C, C)$ means that $\beta_i{<}T$, which allows me to conclude that i chooses to not cooperate out of hate (and regardless of beliefs). I can also conclude that non-cooperators for whom $u(N, C){<}u(C, C)$ needed fear to choose to not cooperate—that is, they chose to not cooperate out of fear.

I find that 24% of the people who do not cooperate do so out of hate, while 76% do so out of fear. Comparing these numbers to those found using the method from Section 3.1 (30% hate, 70% fear) provides a good cross validation exercise. It is reassuring to see that the proportions of people not cooperating out of hate and fear estimated by these two approaches are remarkably similar. Notice it was not clear *a priori* this was going to be the case, as one approach calibrates individual-level parameters out of 8 money allocation decisions, while the other uses everyone's full set of decisions to estimate the distribution of parameters, and then uses an individual's decisions to determine where in the estimated distribution the individual's parameter is likely to be. The fact that the resulting numbers from both approaches are so similar suggests that the results are not the artifact of a particular specification or estimation method.

Counterfactual analysis

I now turn to counterfactual analysis to study how hypothetical policy interventions that shift the drivers of conflict would affect cooperation. First, I investigate how would cooperation change if a policy solved unwarranted fears by correcting misperceptions about the outgroup. In other words, I investigate how would cooperation change if people had accurate beliefs about the percentage of people in the outgroup that is hateful. To do this, I replace everyone's subjective beliefs on the probability that an outgroup member does not cooperate out of hate with the empirical probability of this event happening—that is, I replace $\tilde{P}_i(\beta_j{<}T)$ with $P(\beta_j{<}T)$ for all i. Notice that I can calculate $P(\beta_j{<}T)$ because I have estimated the social preferences of all individuals in both groups. Then, I calculate the expected utilities with the newly imputed beliefs and observe in this scenario how many people prefer to not cooperate, $W(N) > W(C)$. I find that if a policy solved unwarranted fears by correcting inaccurate beliefs about the outgroup, the number of people not cooperating would drop by 73%. This means that 96% of the people who do not cooperate out of fear do so due to misperceptions (recall that 76% of people do not cooperate out of fear, so 73%/76%=96%). This result underlines how misperceptions leading to unwarranted fears are the single most important barrier to intergroup cooperation, and how policies should focus on tackling misperceptions to maximize their effectiveness.

I then investigate how would cooperation change if a policy were to reduce hate. I first simulate a policy that completely eradicates intergroup hate, such that nobody wants to not cooperate out of hate. To do this, I replace the social preferences of all hateful people (those with $\beta_i{<}0$) with selfish social preferences ($\beta_i{=}0$). Then, I calculate the expected utilities with the newly imputed preferences and

observe in this scenario how many people prefer to not cooperate, $W(N) > W(C)$. I find that if a policy completely eradicated intergroup hate, the number of people not cooperating would drop by only 5%. This means such a policy would only manage to convince 21% of the people not cooperating out of hate to switch to cooperation (recall that 24% of people do not cooperate out of hate, so 5%/24%=21%). The reason why the effect is so small is that hateful individuals tend to also be very fearful, so even without hate, most of them will still want to not cooperate out of fear. Interestingly, when I simulate a policy that increases altruism by half a standard deviation on altruistic people (β_i+0.22 for those with β_i>0), the number of people not cooperating drops by 7%. This drop happens because for a higher β_i the level of fear needed to justify not cooperating is higher too (as described by equation (2)). This means that policies that can increase social preferences would be more effective in increasing cooperation if they focused on increasing altruism on altruistic people with fear, than on reducing hate on hateful people. These results highlight how, in this context, policies that are effective at changing social preferences would be ineffective in increasing cooperation, especially if they only targeted hateful people.

The main policy recommendation that can be drawn from these results is that policies that reduce fear would be significantly more effective in increasing cooperation than policies that reduce hate. This is not only because policies to reduce hate affect a smaller percentage of non-cooperators (24% vs 76%), but also because they manage to switch to cooperation a smaller percentage of the population being they target (21% vs 96%).

Data supporting the counterfactual analysis

I now look at the endline data to use within person changes to assess the validity of the counterfactual analysis. I test for the following three lessons. First, to increase cooperation, changing beliefs should be more important than changing preferences. Second, if only preferences are changed, the effects on cooperation should be very small. Third, if only beliefs change, the effects on cooperation should still be considerable. To test these, I run the regression below to study how changes in preferences and beliefs between baseline and endline affected changes in cooperation decisions.

$$\Delta \textit{Cooperate}_i = \phi_0 + \phi_1 \Delta \textit{Hate}_i^{std} + \phi_2 \Delta \textit{Fear}_i^{std} + \varepsilon_i$$

Here, Δ*Cooperate*$_i$ is an indicator variable equal to 1 if i switched to cooperate by endline, 0 if there was no change, and -1 if i switched to not cooperate. $\Delta \textit{Hate}_i^{std}$ is a standardized variable (mean 0, s.d. 1) of the change in negative social preferences i had between the baseline and endline surveys (that is, the change in $-\beta_i$). $\Delta \textit{Fear}_i^{std}$ is a standardized variable (mean 0, s.d. 1) of the change between surveys i had in her beliefs on the probability that the j will not cooperate out of hate (that is, the change in $\tilde{P}_i(\beta_j<T)$). Importantly, because the two regressors are standardized, their coefficients can be compared to determine their relative importance. Note, however, that no exogenous variation is being considered here, so the results are just correlational. Table 2 reports the results.

Table 2: Testing Lessons from the Counterfactual Analysis

	ΔCooperation		
	(1)	(2)	(3)
ΔHate	-0.105***	0.004	
	(0.013)	(0.003)	
ΔFear	-0.191***		-0.184***
	(0.013)		(0.016)
Only if ΔHate=0	N	N	Y
Only if ΔFear=0	N	Y	N
Mean Dep.Var.	.113	.113	.113
Observations	947	316	787

Notes: This table reports correlational effects of the change in preferences and beliefs between the baseline and the endline survey on the change in cooperation. Δ*Cooperate* is an indicator variable equal to 1 if i switched to cooperate by endline, 0 if there was no change, and -1 if i switched to not cooperate. $\Delta Hate_i^{std}$ is a standardized variable (mean 0, s.d. 1) of the change in negative social preferences i had between the baseline and endline surveys (that is, the change in $-\beta_i$). $\Delta Fear_i^{std}$ is a standardized variable (mean 0, s.d. 1) of the change between surveys i had in her beliefs on the probability that the j will not cooperate out of hate (that is, the change in $\tilde{P}_i(\beta_j<T)$). Column 1 includes all subjects. Column 2 restricts the sample to only those subjects who showed no change in fear between surveys. Column 3 restricts the sample to only those subjects who showed no change in hate between surveys. Standard errors are clustered at the individual level.

Column 1 reports that changes in both hate and fear mattered in the decision to switch to cooperation by endline. Crucially, the coefficients also highlight how changing fear was twice as important as changing hate for a switch to cooperation. A decrease in hate by one standard deviation increases the probability of switching to cooperation by around 10 percentage points, while a decrease in fear by one standard deviation increases the probability of switching to cooperation by around 20 percentage points the probability of switching to cooperation.

I then look at the cases where only hate changed or where only fear changed. Column 2 shows that for the cases where only hate changed but fear remained the same, a decrease in hate did not translate into an increase in cooperation. Together with Column 1, these results suggest that changes in hate can increase cooperation only if accompanied by changes in fear. On the other hand, Column 3 shows that for the cases where only fear changed but hate did not, a decrease in fear did translate into an increase in cooperation. Furthermore, the effect in Column 3 is very similar in magnitude to the one found in Column 1. Hence, the results reported in Table 2 bring further evidence to support the lessons drawn from the counterfactual analysis.

Hate, fear and policies of segregation

I now report how my measures of hate and fear correlate with attitudes toward policies against religious segregation in Jos. In a survey module, I asked participants about real integration policies that were that were being discussed by the city and state governments at the time. Specifically, I inquired about a policy promoting integration in settlements and another promoting integration in schools. For each policy, my

Policy	Hateful reason against	Fearful reason against
New settlements in Jos should mix Christians and Muslims	Christians and Muslims have different ways of living that simply cannot coexist together	Some families would not be able to trust their neighbors in these mixed settlements
Schools in Jos should have a mix of Christian and Muslim children and teachers	Muslims and Christians have different ways of educating their children that simply cannot be integrated	The safety of our children would be at risk in these mixed schools

approach to elicit attitudes was structured as follows. I, first, introduced the policy to the participants and stated it "may have some possible downsides." I then presented two potential downsides and asked participants to express the extent to which they agreed or disagreed that each of these downsides was indeed associated with the policy in question. Importantly, the one downside was meant to capture hateful reasons against the policy, while the other was meant to capture fearful reasons against the policy. Participants expressed their level of agreement on a 1 to 4 scale (1=completely disagree, 2=somewhat disagree, 3=somewhat agree, 4=completely agree). The policies and associated downside in this survey module are detailed in the table below.

Table 3: Attitudes on Segregation Policies

	Do you agree the following is a downside of integration in...			
	Settlements		Schools	
	Hateful reason (1)	Fearful reason (2)	Hateful reason (3)	Fearful reason (4)
Hate ($-\beta_i$)	0.488***	0.185**	0.216**	0.164*
	(0.085)	(0.079)	(0.103)	(0.091)
Fear ($\tilde{P}_i(\beta_j<T)$)	0.942***	1.458***	0.672***	1.003***
	(0.168)	(0.144)	(0.171)	(0.143)
Controls	Y	Y	Y	Y
Mean Dep.Var.	1.957	2.397	1.929	1.642
Observations	997	995	996	995

Notes: This table reports the result of regressing the level of agreement with a reason against an integration policy on the lab measures of hate and fear. The outcome variable is a variable from 1 to 4 that expresses how much a person agrees that the stated potential downside of a policy is in fact associated with that policy. The outcome variables in Columns 1 and 2 are about downsides regarding a policy for integration in settlements, while the ones in Columns 3 and 4 are about downsides regarding integration in schools. Columns 1 and 3 are hateful reasons to oppose the policy, while 2 and 4 are fearful reasons to oppose the policy. The variable *Hate* is the negative of the social preferences, $-\beta_i$, and the variable *Fear* is the beliefs on the proportion of the outgroup that is hateful beyond the threshold, $\tilde{P}_i(\beta_j<T)$. Controls are religion, sex and age. Standard errors are clustered at the individual level.

Table 3 reports the results of regressing the level of agreement with a reason against an integration policy on the lab measures of hate and fear. The first thing to note is that the lab measures of hate and

fear exhibit a positive and statistically significant correlation with opposition to two different integration policies for two distinct reasons. This result suggests that my measures effectively capture negative behavior beyond the lab setting. The second thing to note is the strength of the correlation between my hate and fear measures and the reasons for opposing integration. Specifically, my measure of hate displays a stronger association with reasons rooted in hatred against integration, whereas my measure of fear shows a stronger correlation with reasons grounded in fear against integration. To see this, notice the *Hate* coefficient is notably greater for reasons related to hatred than for those related to fear in both policies, while the *Fear* coefficient is notably greater for reasons based on fear compared to those based on hatred in both policies. This evidence suggests that the lab measures used in this paper do not merely capture generic negative attitudes but, rather, discern between attitudes underpinned by hatred and those driven by fear.

6.3. Effects of the radio show

The main specification to study the effects of the radio show is the following.

$$Y_i = \gamma_0 + \gamma_1 \textit{Treated}_i + \gamma_2 X_i + \varepsilon_i$$

Where Y_i is an outcome variable; $\textit{Treated}_i$ is a dummy variable equal to 1 if i belonged to the treatment group; and X_i is a vector of pre-registered controls that includes the outcome variable at baseline, plus other characteristics like religion, sex and age.

Table 4 reports the main effects of the radio show on hate, fear and cooperation. Columns 1 and 2 report the effects on hate. The outcome variable in these columns is the negative of the social preferences for the outgroup, $-\beta_i$, such that a negative coefficient represents a reduction in hate. Columns 3 and 4 report the effects on fear, or the beliefs about the percentage of the outgroup that would want to not cooperate out of hate, $\tilde{P}_i(\beta_j{<}T)$. Columns 5 and 6 report the the effect on the decision to not cooperate in the coordination game.

Table 4A reports the results for the full sample. Columns 1 and 2 show that the radio show reduced hate, although the effect is small in magnitude. Columns 3 and 4 indicate that the radio show had no effect on fear. However, it is worth noting that the point estimates go in the right direction, towards reducing negative beliefs. Columns 5 and 6 show that the show had no effects on cooperation either, although the point estimate has the expected sign.

It is important to note that this first specification might be underestimating the effects of the radio show because it estimates the effects over the full sample, where there are many subjects who are mechanically unresponsive to the treatment because their outcome variable cannot improve from baseline. In other words, many subjects were fully altruistic ($\beta_i{=}1$) or had fully optimistic beliefs ($\tilde{P}_i(\beta_j{<}T) = 0$) at baseline, and therefore they would always show an effect equal to zero, at best. These zeros are not informative of the effectiveness of the policy. Because of this, I run the same regressions restricting

Table 4: Main Effects

A. Full sample

	Hate $-\beta_i$		Fear $\tilde{P}_i(\beta_j<T)$		Non-Cooperation $s_i=N$	
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	-0.026**	-0.026**	-0.012	-0.012	-0.012	-0.014
	(0.013)	(0.012)	(0.011)	(0.010)	(0.015)	(0.014)
Controls	N	Y	N	Y	N	Y
Mean Dep.Var.	-.823	-.823	.218	.218	.169	.169
Observations	947	947	947	947	947	947

B. Removing subjects that are mechanically unresponsive

	Hate $-\beta_i$		Fear $\tilde{P}_i(\beta_j<T)$		Non-Cooperation $s_i=N$	
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	-0.172**	-0.190***	-0.021	-0.021	-0.086	-0.073
	(0.067)	(0.065)	(0.015)	(0.014)	(0.067)	(0.068)
Controls	N	Y	N	Y	N	Y
Mean Dep.Var.	-.079	-.079	.343	.343	1	1
Observations	138	138	600	600	160	160

Notes: This table reports the treatment effect of the radio show. $-\beta_i$ is negative of the social preferences for the outgroup estimated following the approach presented in Section 3.2. $\tilde{P}_i(\beta_j<T)$ is the beliefs on the percentage of the outgroup that will not cooperate out of hate. $s_i=N$ is the decision to not cooperate in the coordination game. The controls are the outcome variable at baseline, religion, sex and age. Table 1A report results for the full sample. Table 2A restricts the sample to individuals who were not mechanically unresponsive in the outcome variable of the respective column. Standard errors are clustered at the individual level.

the sample to individuals who had margin for improvement in the outcome variable of the respective column. Results are reported in Table 4B. Columns 1 and 2, show there was a reduction in hate that is considerably greater than the one previously estimated. In particular, Column 2 indicates that listening to the radio show reduced hate by 0.19 units for this groups, which is 0.45 of a standard deviation. Columns 3 and 4 still show there were no effects on fear, although the point estimates doubles with respect to the previous estimation. And Columns 5 and 6 show that there is still no effect on cooperation, although the point estimates increases substantially.

Finding that the radio show reduces hate but does not increase cooperation could have been a puzzling result that would make it difficult to conclude if the policy was ultimately effective or not. However, this result becomes easy to rationalize given the analysis done in Section 6.2. The radio show is an effective policy because it reduces hate, but it is the wrong policy for this context because it does not affect the key motive for conflict, which is fear. This ultimately renders the policy ineffective at achieving its main goal, which is increasing cooperation.

Figure 3 can also help illustrate this result. Consider the case of individuals with preferences β_i=−.25, who have hateful preferences just above the threshold. And notice that these participants believe that at least 40% of the outgroup is hateful. The radio show moved the dots of these individuals to the right, leaving them close to β_i=−.05, and effectively removing their hateful motive to not cooperate. However, the radio show does not affect beliefs, so these dots do not move vertically. Importantly, the theory presented in Section 3 states that with β_i=−.05, it is enough to believe that 23% of the outgroup is hateful to want to not cooperate. Therefore, these participants will still want to not cooperate out of fear, based on this simple graphical counterfactual.

This result highlights the perils of designing policy interventions without understanding the drivers of non-cooperation.

Heterogeneous effects by baseline level of hate and fear

Table 5: Heterogeneous effects

	Hate	Fear	Non-Cooperation	
	$-\beta_i$	$\tilde{P}_i(\beta_j<T)$	$s_i=N$	
	(1)	(2)	(3)	(4)
Treated x Hate$_{t=0}$	-0.093***		0.003	
	(0.033)		(0.029)	
Treated x Fear$_{t=0}$		-0.024*		-0.018
		(0.013)		(0.022)
Treated	-0.026**	-0.012	-0.012	-0.012
	(0.012)	(0.010)	(0.014)	(0.014)
Controls	Y	Y	Y	Y
Mean Dep.Var.	-.823	.218	.169	.169
Observations	947	947	947	947

Notes: This table reports the heterogenous treatment effect of the radio show. $-\beta_i$ is negative of the social preferences for the outgroup estimated following the approach presented in Section 3.2. $\tilde{P}_i(\beta_j<T)$ is the beliefs on the percentage of the outgroup that will not cooperate out of hate. $s_i=N$ is the decision to not cooperate in the coordination game. Hate$_{t=0}$ and Fear$_{t=0}$ refer to the outcome variables of Column 1 and 2 at baseline. The controls are the outcome variable at baseline, religion, sex and age. Standard errors are clustered at the individual level.

I now look at how treatment effects varied depending on how hateful or fearful participants were at baseline. To do this, I run the same specification as before and add as a regressor the interaction between the treatment variable and preferences or beliefs at baseline. Results are reported in Table 5. Regarding social preferences, Column 1 shows that the effects were strongest for the most hateful people. It was not obvious *a priori* that this would be the case, as it was plausible that the most hateful individuals would have more rigid preferences. However, I find that social preferences increase by around 0.12 units in fully hateful people and that this effect decreases progressively until there is none in people with β_i=0.3. It is worth recalling that, according to the counterfactual analysis of Section 6.2, the radio

show would be more effective in increasing cooperation if it increased the social preferences of altruistic people than of hateful people. Therefore, because the effect of the radio show is concentrated on the most hateful people, it was even less likely that this effect would translate into increased cooperation. Regarding beliefs, Column 2 shows some evidence that the radio show did reduce fear, and that this effect was strongest in the people with the greatest misperceptions. Taken at face value, the coefficient says that people who believed that 100% of the outgroup was hateful adjusted their beliefs to 76%. Finally, Columns 3 and 4 look at how the heterogeneous effects on hate and fear could have deferentially affected cooperation. However, I find no evidence that the effects identified in Columns 1 and 2 translated into increased cooperation.

Social desirability bias

One potential threat to the results in this section is that they are driven by social desirability bias. Despite the choices made on the experimental design to reduce demand effects, one might be concerned that the treatment group could express more social desirability bias than the control group. Participants who listened to a radio show that aimed to promote intergroup cooperation might disingenuously express more positive attitudes towards the outgroup to present themselves in a good light to the surveyors. I now show evidence that this was not the case. Following Dhar, Jain & Jayachandran (2022), I include in the baseline survey a module (developed by psychologist) that measures a person's propensity to give socially desirable answers. The module asks respondents if they have several too-good-to-be-true traits such as never being jealous, lazy or resentful. Those who report having more of these traits are scored as having a higher propensity to give socially desirable answers. I use these individual-level scores to see if subjects with a higher propensity to have social desirability bias seem to be more positively affected by the radio show (which could drive the results). To test for this I run the following regression.

$$Y_i = \eta_0 + \eta_1 Treated_i + \eta_2 SDS_i + \eta_3 Treated_i \times SDS_i + \eta_4 X_i + \varepsilon_i$$

Where Y_i is an outcome variable; $Treated_i$ is a dummy variable equal to 1 if i belongs to the treatment group; SDS_i is a variable from 1 to 10 indicating how many socially desirable answers i gave in that survey module; and X_i is a vector of controls that includes the outcome variable at baseline, plus other characteristics like religion, sex and age. Table 6 reports the results.

First, Row 3 indicates that participants with a higher tendency to give socially desirable answers indeed expressed less hate and less fear towards the outgroup in the survey. This result is also a validation of the measurement of social desirability provided by the survey module. Crucially, Row 1 shows that the tendency to give more socially desirable answers was not higher in the treatment group vs. the control group, which indicates that listening to the radio show did not increase demand effects. In addition, Column 1, Row 2, shows that the treatment effect on hate is robust to controlling for social desirability, and what is more, this effect is of the exact same magnitude as that reported in the main

Table 6: Social Desirability Bias

	Hate $-\beta_i$	Fear $\tilde{P}_i(\beta_j<T)$	Non-Coo. $s_i=N$
	(1)	(2)	(3)
Treated x SDS	0.007	0.007	0.002
	(0.007)	(0.005)	(0.008)
Treated	-0.026**	-0.012	-0.013
	(0.012)	(0.010)	(0.014)
Soc.Des.Score	-0.007*	-0.007*	-0.009
	(0.004)	(0.004)	(0.006)
Controls	Y	Y	Y
Mean Dep.Var.	-.823	.218	.169
Observations	947	947	947

Notes: This table reports the treatment effect of the radio show controling for social desirability bias. $-\beta_i$ is negative of the social preferences for the outgroup estimated following the approach presented in Section 3.2. $\tilde{P}_i(\beta_j<T)$ is the beliefs on the percentage of the outgroup that will not cooperate out of hate. $s_i=N$ is the decision to not cooperate in the coordination game. *Soc.Des.Score* and *SDS* refer to the individual-level social desirability bias score. The controls are the outcome variable at baseline, religion, sex and age. Standard errors are clustered at the individual level.

specification (Table 2A). Taken together, these results provide evidence that the effects of the radio show do not appear to be driven by social desirability bias.

Average treatment effect on the treated

Finally, I address the issue that only 33% of the participants answered at least one of the quizzes about the radio show. It is hard to know to what extent this percentage reflects the number of people who actually listened to the radio show. However, it suggests that it might have been the case that a majority of people did not listen to the radio show. If this is the case, one would like to estimate the treatment effect on the treated (ATT). Under the assumption that everyone who listened to the radio show answered at least one quiz and everyone who did not listened answered no quiz, I can use the treatment assignment as an IV for listening to the radio show, and estimate the ATT. The results of this exercise are reported in the Appendix. I find that the ATT on hate is -0.08 units (3.1 times the ATE), significant at the 5% level. I find no effects on fear or cooperation, consistent with the previous findings.

7. Conclusion

In this paper I develop a theory-driven experimental protocol to disentangle hate and fear. I then use this protocol to understand, first, what drives conflict in a particular setting, and, second, which drivers a particular policy intervention shifts. I find that unwarranted fears are the main driver of conflict and that interventions should focus on solving misperceptions about the outgroup to maximize the impact

on cooperation. However, I also find that unwarranted fears prove hard to change with policy (even more than hate).

How generalizable are these results? One plausible explanation for unwarranted fears towards the outgroup, which may point at a broader pattern, is that "bad" individuals are more salient because their actions receive more media coverage and are more talked about. In line with well-studied psychological biases (like availability bias or 'what you see is all there is' bias (Enke, 2020)), people will not realize that the information they receive exaggerates the number of "bad" individuals in the outgroup, which would lead to widespread misperceptions. In the US, for example, after the 2021 Capitol attack, Democrats greatly exaggerated the number of Republicans supporting political violence (Mernyk et al, 2022).

This paper also suggests that changing beliefs about the outgroup is harder than changing preferences for the outgroup. There is evidence that suggests that this finding could be general. Two different experimental intergroup contact interventions, conducted in India and Iraq, find a similar pattern: individuals in the treatment group increased positive behavior toward the outgroup, but their trust in the outgroup remained unchanged or even decreased (Lowe, 2021; Mousa, 2020). Furthermore, there is evidence that after conflict, broken trust between communities resulting from violence against civilians seems strikingly resilient, lasting for generations (Tur-Prats and Valencia-Caicedo, 2023).

However, further research is needed to assess the extent of these conclusions. Importantly, the experimental protocol in this paper is portable and can be deployed elsewhere. Using this protocol in different settings and for different policies can advance the understanding of intergroup conflict and its solutions.

References

Acemoglu, D. and A. Wolitzky (2023). Mistrust, misperception, and misunderstanding: Imperfect information and conflict dynamics.

Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015, November). Radio and the Rise of The Nazis in Prewar Germany. *The Quarterly Journal of Economics 130*(4), 1885–1939.

Adida, C. L., A. Lo, and M. R. Platas (2018, September). Perspective taking can promote short-term inclusionary behavior toward Syrian refugees. *Proceedings of the National Academy of Sciences 115*(38), 9521–9526. Publisher: Proceedings of the National Academy of Sciences.

Alan, S., C. Baysan, M. Gumren, and E. Kubilay (2021, November). Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking*. *The Quarterly Journal of Economics 136*(4), 2147–2194.

Alesina, A. and E. L. Ferrara (2005, September). Ethnic Diversity and Economic Performance. *Journal of Economic Literature 43*(3), 762–800.

Anderson, S. (2011, January). Caste as an Impediment to Trade. *American Economic Journal: Applied Economics 3*(1), 239–263.

Bauer, M., C. Blattman, J. Chytilová, J. Henrich, E. Miguel, and T. Mitts (2016, September). Can War Foster Cooperation? *Journal of Economic Perspectives 30*(3), 249–274.

Bauer, M., A. Cassar, J. Chytilová, and J. Henrich (2014, January). War's Enduring Effects on the Development of Egalitarian Motivations and In-Group Biases. *Psychological Science 25*(1), 47–57. Publisher: SAGE Publications Inc.

Blattman, C. and E. Miguel (2010, March). Civil War. *Journal of Economic Literature 48*(1), 3–57.

Bonomi, G., N. Gennaioli, and G. Tabellini (2021, November). Identity, Beliefs, and Political Conflict. *The Quarterly Journal of Economics 136*(4), 2371–2411.

Broockman, D. and J. Kalla (2016, April). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science 352*(6282), 220–224. Publisher: American Association for the Advancement of Science.

Bursztyn, L., T. Chaney, T. A. Hassan, and A. Rao (2021, February). The Immigrant Next Door: Long-Term Contact, Generosity, and Prejudice.

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

Casey, K. (2015, August). Crossing Party Lines: The Effects of Information on Redistributive Politics. *American Economic Review 105*(8), 2410–2448.

Charness, G. and Y. Chen (2020). Social Identity, Group Behavior, and Teams. *Annual Review of Economics 12*(1), 691–713. _eprint: https://doi.org/10.1146/annurev-economics-091619-032800.

Chassang, S. and G. Padró i Miquel (2007). Mutual Fear and Civil War.

Chen, Y. and S. X. Li (2009, March). Group Identity and Social Preferences. *American Economic Review 99*(1), 431–457.

Choi, J.-K. and S. Bowles (2007, October). The Coevolution of Parochial Altruism and War. *Science 318*(5850), 636–640. Publisher: American Association for the Advancement of Science.

DellaVigna, S. (2018, January). Structural Behavioral Economics. In B. D. Bernheim, S. DellaVigna, and D. Laibson (Eds.), *Handbook of Behavioral Economics: Applications and Foundations 1*, Volume 1 of *Handbook of Behavioral Economics - Foundations and Applications 1*, pp. 613–723. North-Holland.

DellaVigna, S., R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya (2014, July). Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia. *American Economic Journal: Applied Economics 6*(3), 103–132.

DellaVigna, S. and E. La Ferrara (2015, January). Economic and Social Impacts of the Media. In S. P. Anderson, J. Waldfogel, and D. Strömberg (Eds.), *Handbook of Media Economics*, Volume 1 of *Handbook of Media Economics*, pp. 723–768. North-Holland.

Dhar, D., T. Jain, and S. Jayachandran (2022, March). Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India. *American Economic Review 112*(3), 899–927.

Enke, B., R. Rodríguez-Padilla, and F. Zimmermann (2023, July). Moral Universalism and the Structure of Ideology. *The Review of Economic Studies 90*(4), 1934–1962.

Enos, R. D. (2014, March). Causal effect of intergroup contact on exclusionary attitudes. *Proceedings of the National Academy of Sciences 111*(10), 3699–3704. Publisher: Proceedings of the National Academy of Sciences.

Falk, A. and C. Zehnder (2013, April). A city-wide experiment on trust discrimination. *Journal of Public Economics 100*, 15–27.

Fearon, J. D. and D. D. Laitin (1996, December). Explaining Interethnic Cooperation. *American Political Science Review 90*(4), 715–735. Publisher: Cambridge University Press.

Fershtman, C. and U. Gneezy (2001). Discrimination in a segmented society: An experimental approach. *The Quarterly Journal of Economics 116*(1), 351–377.

Fouka, V. and M. Tabellini (2022, August). Changing In-Group Boundaries: The Effect of Immigration on Race Relations in the United States. *American Political Science Review 116*(3), 968–984. Publisher: Cambridge University Press.

Franck, R. and I. Rainer (2012, May). Does the Leader's Ethnicity Matter? Ethnic Favoritism, Education, and Health in Sub-Saharan Africa. *American Political Science Review 106*(2), 294–325. Publisher: Cambridge University Press.

Francois, P., I. Rainer, and F. Trebbi (2015). How is power shared in africa? *Econometrica 83*(2), 465–503.

Ghosh, A. (2022, August). Religious Divisions and Production Technology: Experimental Evidence from India.

Giuliano, L., D. I. Levine, and J. Leonard (2009, October). Manager Race and the Race of New Hires. *Journal of Labor Economics 27*(4), 589–631. Publisher: The University of Chicago Press.

Hjort, J. (2014, November). Ethnic Divisions and Production in Firms. *The Quarterly Journal of Economics 129*(4), 1899–1946.

Hodler, R. and P. A. Raschky (2014, May). Regional Favoritism. *The Quarterly Journal of Economics 129*(2), 995–1033.

Jha, S. (2013). Trade, institutions, and ethnic tolerance: Evidence from south asia. *American political Science review 107*(4), 806–832.

Korovkin, V. and A. Makarin (2023, January). Conflict and Intergroup Trade: Evidence from the 2014 Russia-Ukraine Crisis. *American Economic Review 113*(1), 34–70.

Kramon, E. and D. N. Posner (2016, April). Ethnic Favoritism in Education in Kenya. *Quarterly Journal of Political Science 11*(1), 1–58. Publisher: Now Publishers, Inc.

Kranton, R., M. Pease, S. Sanders, and S. Huettel (2020, September). Deconstructing bias in social preferences reveals groupy and not-groupy behavior. *Proceedings of the National Academy of Sciences 117*(35), 21185–21193. Publisher: Proceedings of the National Academy of Sciences.

Kromka, S. M. and A. K. Goodboy (2019, January). Classroom storytelling: using instructor narratives to increase student recall, affect, and attention. *Communication Education 68*(1), 20–43. Publisher: Routledge _eprint: https://doi.org/10.1080/03634523.2018.1529330.

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

Lowe, M. (2021, June). Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration. *American Economic Review 111*(6), 1807–1844.

Luttmer, E. F. P. (2001, June). Group Loyalty and the Taste for Redistribution. *Journal of Political Economy 109*(3), 500–528. Publisher: The University of Chicago Press.

Marx, B., V. Pons, and T. Suri (2021, May). Diversity and team performance in a Kenyan organization. *Journal of Public Economics 197*, 104332.

Michelitch, K. (2015, February). Does Electoral Competition Exacerbate Interethnic or Interpartisan Economic Discrimination? Evidence from a Field Experiment in Market Price Bargaining. *American Political Science Review 109*(1), 43–61. Publisher: Cambridge University Press.

Mousa, S. (2020, August). Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq. *Science 369*(6505), 866–870. Publisher: American Association for the Advancement of Science.

Oh, S. (2023, August). Does Identity Affect Labor Supply? *American Economic Review 113*(8), 2055–2083.

Padró i Miquel, G. (2007, October). The Control of Politicians in Divided Societies: The Politics of Fear. *The Review of Economic Studies 74*(4), 1259–1274.

Paluck, E. L. (2009, March). Reducing intergroup prejudice and conflict using the media: a field experiment in Rwanda. *Journal of Personality and Social Psychology 96*(3), 574–587.

Paluck, E. L., S. A. Green, and D. P. Green (2019, November). The contact hypothesis re-evaluated. *Behavioural Public Policy 3*(2), 129–158. Publisher: Cambridge University Press.

Paluck, E. L., R. Porat, C. S. Clark, and D. P. Green (2021). Prejudice Reduction: Progress and Challenges. *Annual Review of Psychology 72*(1), 533–560. _eprint: https://doi.org/10.1146/annurev-psych-071620-030619.

Rabin, M. (2000). Risk Aversion and Expected-Utility Theory: A Calibration Theorem. *Econometrica 68*(5), 1281–1292. Publisher: [Wiley, Econometric Society].

Rao, G. (2019, March). Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools. *American Economic Review 109*(3), 774–809.

Rubinstein, A. (1989). The Electronic Mail Game: Strategic Behavior Under "Almost Common Knowledge". *The American Economic Review 79*(3), 385–391. Publisher: American Economic Association.

Scacco, A. and S. S. Warren (2018, August). Can Social Contact Reduce Prejudice and Discrimination? Evidence from a Field Experiment in Nigeria. *American Political Science Review 112*(3), 654–677. Publisher: Cambridge University Press.

Shayo, M. (2020). Social Identity and Economic Policy. *Annual Review of Economics 12*(1), 355–389. _eprint: https://doi.org/10.1146/annurev-economics-082019-110313.

Slater, M. D. and D. Rouner (2002, May). Entertainment-Education and Elaboration Likelihood: Understanding the Processing of Narrative Persuasion. *Communication Theory 12*(2), 173–191.

Trebbi, F. and E. Weese (2019). Insurgency and small wars: Estimation of unobserved coalition structures. *Econometrica 87*(2), 463–496.

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

Appendix

You can find the Appendix here.