UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Hate, Fear and Intergroup Conflict

Permalink

Author

Ortiz, Miguel

Publication Date

2024

Peer reviewed|Thesis/dissertation

Hate, Fear and Intergroup Conflict

By

Miguel Ortiz

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Ernesto Dal Bó, Co-chair
Francesco Trebbi, Co-chair
Frederico Finan
Edward Miguel

Spring 2024

Hate, Fear and Intergroup Conflict

Abstract

Hate, Fear and Intergroup Conflict

by

Miguel Ortiz

Doctor of Philosophy in Business Administration

University of California, Berkeley

Ernesto Dal Bó, Co-chair

Francesco Trebbi, Co-chair

Understanding the deep drivers of social conflict is crucial for identifying policies that can effectively reduce it. This dissertation studies how intergroup conflict can be driven by hate and fear, and how policy interventions can affect these drivers to increase cooperation between groups. In Chapter 1 I develop a theoretical model of conflict, with hate and fear as primitives. In Chapter 2, I use this model to develop an experimental protocol to empirically estimate the extent to which intergroup conflict is driven by hate vs. fear. I deploy this protocol as a lab-in-the-field experiment in 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 between Christians and Muslims. Moreover, this fear is mostly unwarranted, as non-cooperators grossly exaggerate the percentage of hateful people in the outgroup. In Chapter 3, I turn to policy and study why policies currently trusted to promote intergroup cooperation may or may not be effective at doing so. To this end, I do 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 that this decrease in fear does not translate into increased cooperation.

Contents

Acknowledgments

I am especially grateful to Ernesto Dal Bó and Francesco Trebbi for their generous support and guidance throughout these years. I was lucky to find two advisors who allowed me to think creatively and helped me elevate my work beyond my expectations. 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.

The PhD would be an empty experience without friends—you were the best part of my PhD and I am deeply grateful for all of you. To my housemates, Felipe Arteaga, Andrea Cerrato and Matteo Tranchero, every dinner I felt lucky for being a part of the home we built. To Silvia Barbereschi, Javier Feinmann, Roberto Hsu Rocha and David Qihang Wu, I'll cherish your support and all the happy moments we shared together. To Rafael Unda, you were my family in Berkeley.

I am also profoundly grateful to my family. To my parents, who are a stellar example of living *the good life*—both in enjoyment and virtue. And my sisters, Laura, Sofía, María and Juana, each one of you makes me feel so loved, and that is something I value everyday.

I also want to thank UC Berkeley and the Business and Public Policy program at the Haas School of Business, they are amazing institutions of which I feel proud of having been a part of. And last but not least, I am grateful to the Berkeley/Oakland/San Francisco area. I feel very lucky to have lived in this place for six years. All the experiences this place has to offer left me with countless good memories, and the amazing culture of this place influenced me and is now a part of who I am.

Introduction

Societies are often fractured into social identity groups, like ethnicities or religions, which can result in substantial economic losses (Easterly and Levine, 1997; Alesina and La Ferrara, 2005). Fractionalization and polarization give rise to "ingroup" vs. "outgroup" dynamics that are usually characterized by hostile behavior and social conflict (Tajfel and Turner, 2004). These hostilities can take many forms, encompassing labor discrimination, the curtailment of civil rights and access to public goods, segregation, and even war and genocide. Despite this being a recurrent pattern in history and across the globe, we still lack clarity on what motives drive citizens to support hostile actions against the outgroup and fuel social conflict.

Studying the individual-level drivers of conflict is hard, because there are many potential motives, few with a clear definition that distinctly separates them from the others, and even fewer with an empirical measure. To make progress, I focus on the fundamental partition of motives for action that is preferences vs. beliefs. Support for hostilities may be rooted in preferences: individuals harbor hatred towards the outgroup and want to harm them. Alternatively, support for hostile behavior 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 of policy interventions to promote intergroup cooperation. Nevertheless, hostile behavior is an equilibrium outcome, making it empirically challenging to disentangle preferences from beliefs.

This dissertations studies how intergroup conflict cab be driven by hate and fear, and how policy interventions can affect these channels to promote cooperation between groups.

In **Chapter 1**, I develop a theoretical model of conflict, with hate and fear as primitives. 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). I model conflict between groups as a coordination game, where cooperation is a Nash equilibrium that 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 the payoff of the outgroup player by a greater amount. Alternatively, non-cooperation may stem also from fear. A player who is not hateful but fears the outgroup player may be hateful (and therefore non-cooperative) will also want to not cooperate. The model, therefore, explicitly distinguishes between the roles of preferences and beliefs in driving non-cooperation.

In **Chapter 2**, I seek to empirically answer to what extent is intergroup conflict driven by hate vs. fear? I further inquire: If conflict is driven by fear, is this fear warranted? In other words, are beliefs about the preferences of the outgroup accurate, or are they misperceptions? To answer this question, I use the model developed in Chapter 1 to guide the design of an experimental protocol to empirically disentangle the preferences and beliefs that drive hostile action and assess their relative importance. Then, 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. Jos is a city where Christians and Muslims have lived together for over a hundred years. In the past 20 years, the city experienced multiple outbreaks of religious violence perpetrated by ordinary citizens from both sides. These events 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.

In the field, I measure cooperation between Christians and Muslims through coordination games. However, 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 outcomes, such that beliefs do not enter the decision 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 I ask participants to 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 the 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 identify their beliefs about the outgroup members' level of hate (or altruism). I then use the information collected to estimate a structural model to recover social preferences at the individual level. Using the estimated preferences and elicited beliefs, I am able to 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 levels in this context. 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 mutual cooperation is an equilibrium and yields 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 payoffs 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 individual decisions made in the game. The estimated model leads to three main findings. First, 24% of non-cooperation decisions are motivated by hate, while 76% are 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. This difference in behavior between religious groups is most likely driven by the salience of the armed group Boko-Haram in this area, which generates disproportionately negative feelings toward one of the groups.

Having understood what drives intergroup conflict, in **Chapter 3** I turn the focus to policy an ask: Are local policies currently trusted to promote cooperation able to address what I identify to be the key driver of conflict? Here I focus on one particular type of policy which is cultural media interventions (CMIs). CMIs are a popular intervention in Africa to tackle different social issues (Banerjee et al., 2020) and have received increased attention as a policy to mitigate conflict (e.g., in Rwanda (Paluck, 2009)). In particular, this chapter

studies radio drama series, a prime example of CMIs.

To study why CMIs may or may not be effective at increasing cooperation, 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 to mitigate 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 or too costly. 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. The 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 inefficient at achieving its main goal (increasing cooperation), because it does not affect the key driver of non-cooperation, which is fear. 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 also 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 main result is not likely driven by social desirability bias, following the methodology in Dhar et al. (2022).

Taken together, the results of this dissertation illustrate the value of the model-based protocol to explain conflict behavior and policy efficacy. The protocol serves a dual purpose: first, it can diagnose what drives conflict in a particular setting, and second, it can identify which drivers of conflict a specific policy can shift. Furthermore, by comparing these two results, the protocol allows to assess the alignment or misalignment between a place's needs

and a policy. For the case of this dissertation, 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 to study other settings and policies can advance the understanding of intergroup conflict and its solutions.

This dissertation makes three main contributions. First, it presents a 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 dissertation builds upon the findings and tools of that literature to develop a theory of intergroup conflict and a protocol to estimate the parameters of the theory.

Second, this dissertation provides empirical evidence on how unwarranted fears can be an essential driver of conflict and how policies may struggle to change them. This finding directly contributes to the theoretical literature on fear and conflict (Chassang and Padró i Miquel, 2007; Acemoglu and Wolitzky, 2023) by providing well-identified evidence that fears are an empirically relevant driver of conflict. 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. 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; Dal Bó and Dal Bó, 2014). Furthermore, because I use a general framework of cooperation, the findings could help explain what drives inefficient behavior in multiple other spheres of intergroup interaction. Inefficient intergroup interactions have been documented in economic performance (Hjort, 2014; Ghosh, 2022; Marx et al., 2021; Alesina and La Ferrara, 2005), labor selection (Oh, 2023; Giuliano et al., 2009), 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 dissertation provides experimental evidence on the effectiveness of using cultural media 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), 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 dissertation sheds light on these mechanisms. In addition, my 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 media influences attitudes of people in conflict situations (Yanagizawa-Drott, 2014; Adena et al., 2015; DellaVigna et al., 2014; Paluck, 2009).

Chapter 1

A Model on the Deep Drivers of Conflict

In this chapter I develop a theoretical model of conflict, where hate and fear are primitives of hostile behavior. The model explicitly distinguishes between the roles of preferences and beliefs in driving non-cooperation.

1.1. The base model

Consider a society with two groups, I (the ingroup) and O (the outgroup). Let i be a member of group I, and j a member of group O. When interacting with j, i has the the following utility function:

$$u_i = x_i + \beta_i(Z_i) \cdot x_j \tag{1.1}$$

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 O. If $\beta_i < 0$, i is hateful towards members of group O; if $\beta_i > 0$, i is altruistic towards members of group O; 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 indicates a vector past experiences, education, individual characteristics, etc. In what follows, I take Z_i as predetermined, but an alternative model can be found in the Appendix, where β_i is endogenous and depends on the beliefs about β_j. In addition, the Appendix contains a model extension where $j \in \{I, O\}$.

Members of the different groups face each other in a coordination game (or stag-hunt game), where there are two strategies: *Cooperate* (C) or *Not Cooperate* (N)[1]. For directness,

---

[1]Coordination games can allow for a richer study of the reasons behind non-cooperation, compared

I illustrate the theory with the following coordination game, which is the one subjects face in the experiment (where payoff units are in Nigerian naira). Payoffs in the matrix represent x_i and x_j in the utility function.

Example 1

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

In the case where there are no social preferences, $\beta_i{=}0$, 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, $u_i(N, N) > u_i(C, N)$.[2] Intuitively, even if i is fully altruistic and has $\beta_i{=}1$, 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 $u_i(N, C) > u_i(C, C)$. In our example,

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

Define the threshold $T{\equiv}-0.2$. If i is hateful beyond the threshold, she will prefer to not cooperate regardless of what j will do. That is, if $\beta_i{<}T$, N is a dominant strategy for i. In the case of Example 1, $\beta_i{<}T$ means that i is hateful enough to prefer to lose 100 and reduce j's payoff by 500. When $\beta_i{<}T$, 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

---

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.

[2]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 $\beta_i{<}5/3$.

($\beta_j{<}T$), 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 $\beta_j{<}T$, 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 O. 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:

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

Given this, i chooses to not cooperate if $W_i(N) \geq W_i(C)$. In the case of Example 1, solving for $\tilde{P}_i(s_j{=}N)$ yields the following.

$$\tilde{P}_i(s_j{=}N) \geq \frac{2}{7}\left(\frac{1+5\beta_i}{1+\beta_i}\right) \tag{1.2}$$

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; Kneeland, 2015), in the empirical model I will assume that players do not form higher-order beliefs when playing the game— i.e., they are limited to $k{=}1$ level reasoning. 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 Appendix B2 I provide evidence that participants in the experiment do not seem to form higher-order beliefs, and that preferences and first order beliefs explain almost all the variation in the cooperation decisions. Therefore, this assumption appears to be the best representation of behavior.

1.2. Model extensions

Endogenous social preferences

An individual's base utility function is

$$u_i = x_i + \beta_i(Z_i){\cdot}x_j \tag{1.3}$$

In the model presented previously Z_i is said to be predetermined. The following extension of the model considers the case in which social preferences are endogenous to the beliefs about the social preferences of others (in the spirit of Rabin (1993) and Levine (1999)). This formulation allows people to have social preferences that depend on reciprocity. In other words, people may have higher social preferences for those who they belief have higher social preferences towards them, and lower social preferences for those who they believe have lower social preferences towards them.

Let $\tilde{\beta}_j$ be i's believe about the expected social preferences of j given that j belongs to the ougroup O—that is, i's beliefs about $E[\beta_j|j\in O]$. Then, $\beta_i(\tilde{\beta}_j)$. In particular, β_i has the following functional form.

$$\beta_i = \frac{\alpha_i + \lambda\tilde{\beta}_j}{1 + \lambda}$$

$$u_i = x_i + \left(\frac{\alpha_i + \lambda\tilde{\beta}_j}{1 + \lambda}\right) \cdot x_j$$

Where $\alpha_i\in[-1, 1]$ is the base social preferences of i. The base social preferences are adjusted by a reciprocity parameter $\lambda\in[0, 1]$. A $\lambda > 0$ means that i wants to adjust her base social preferences in order to correspond to the social preferences she believes j has towards her. Note that in this formulation it is still the case that $\beta_i\in[-1, 1]$. And note that when $\lambda=0$ and there is no reciprocity in social preferences, $\beta_i=\alpha_i$ and the model goes back to its original formulation written in this chapter.

Utility function with different social preferences parameters for the ingroup and the outgroup

In the model presented previously it is always the case that j belongs to the outgroup, $j\in O$. One simple extension allows for j to belong to the outgroup or the ingroup, $j\in\{I, O\}$, and for i to have different social preferences for j depending on which group j belongs to. In this case, i has one social preference parameter for people from the in group, β_{iI}, and one different social preference parameter for people from the outgroup, β_{iO}. The utility function is the following

$$u_i(x_i, x_j) = x_i + \big(\beta_{iI} \cdot \mathbb{1}(j\in I) + \beta_{iO} \cdot \mathbb{1}(j\in O)\big)x_j$$

With this formulation, depending on which group j belongs to, i uses a different social preference parameter in the interaction. Additionally, the ratio $\frac{\beta_{iI}}{\beta_{iO}}$ captures i's level of

ingroup bias (or moral universalism), where $\frac{\beta_{iI}}{\beta_{iO}}=0$ means i has no ingroup bias (or full moral universalism), and higher a $\frac{\beta_{iI}}{\beta_{iO}}$ means a higher level of ingroup bias.

Chapter 2

An Empirical Study of Hate and Fear as Drivers of Conflict

In this chapter I use the theory presented previously to guide the design of an experimental protocol to empirically disentangle hate and fear and determine their relative importance. I this chapter I also expose the results I find when I deploy this experimental protocol in Nigeria to study the conflict between Christians and Muslims in the region.

2.1. Lab experiment design

This section presents the experimental protocol that identifies 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 outcomes, such that beliefs do not enter the decision 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 of preferences vs. beliefs, consider the following example.

Example 2

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

Coordination Game

	You	Other
Opt 1	1000	1000
Opt 2	900	500

Money Allocation Decision

↑   ←—   ↑

preferences   preferences

+ beliefs

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 categorize people into two groups by examining their decisions in the two situations above together. 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 presents to the participant exactly 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.[1] As I explain below, in the experiment participants face multiple money allocation decisions, which allows me to determine with more precision where their social preferences lie. Below I also explain how I ask participants to guess others' money allocation decisions to elicit their beliefs about others' social preferences.

Experiment set up

With the insight from above in mind, the lab experiment proceeds as follows. First, participants are told there are two groups, the Blue Group and the Green Group, and that they will play with one group first and then the other. Before playing with each group, they are shown the list of names of all the members of the group, and are told that they will be randomly matched and play with one of the members of the group but won't know which one. Each group consists of ten individuals who made their decisions in advance. 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

---

[1]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.

can easily identify that the common characteristic of the members of each group is religion. After playing a set of 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 played the same set of games with each group.

This design has two advantages. First, using names to signal religion allows me to not mention religion explicitly, which helps reduce experimenter demand bias[2]. Second, by not knowing exactly 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 signal older or younger cohorts.

Importantly, there was no deception in this experiment. The names in the Blue and Green groups belonged to real people who were sampled from the pilot of the experiment and made their decisions in advance. Payoffs in the games were implemented. In addition, the names of participants (that were not from the pilot) were never recorded, to lower demand bias.

Eliciting social preferences

After participants are matched with an anonymous person from the first group shown to them, they start the activities of the experiment with a series of money allocation decisions. There are 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, of which the payoffs of only one are implemented (picked at random). 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, at a cost for the decision-maker.[3] In the case of the money allocation decision in Example 2 above, to cost is 1000-900=100. In the other half of the 20 money allocation decisions, Option 2 represents the altruistic option of increasing the match's payoff in ₦500 naira, at a cost for the decision-maker. 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 elicits 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

---

[2]In the results section I also check how robust is the analysis to controlling for social desirability bias.
[3]₦500 naira represents half the hourly wage in this context.

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.[4] 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, it ignores sampling variability, so I cannot calculate standard errors for the individual-level parameters I have calibrated. 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 Section 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.

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 an 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 same series of money allocation decisions as before, but this time trying to guess what their match from the Green/Blue group picked. This elicits the beliefs about the mean of the distribution, $\tilde{E}_i[\beta_j]$. Afterward, I elicit beliefs on the mass of the tail of the distribution after threshold T in the following way. I ask participants to guess how many people (out of the 10) from the Green/Blue group picked Option 2 in one particular money allocation decisions. This scenario represents the threshold where someone would reveal to be hateful enough to want to not cooperate out of hate in an upcoming coordination game. In other words, this money allocation decisions mirrors an upcoming coordination game as showcased in Example 2. 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 $\tilde{P}_i(s_j = N)$. Because participants play two coordination games (explained below), this previous exercise is done twice (once for each money allocation decision that mirrors each game).

---

[4]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 $\beta_i \leq (x_{i2} - x_{i1})/(x_{j1} - x_{j2})$.

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 variation in payoffs that changes the threshold of how hateful a person needs to be to want to not cooperate out of hate. For the first game the threshold is $T=-0.2$, and for the second one the threshold is $T=-0.6$. After these, the activities of the lab conclude. Full details on the experimental 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.

2.2. 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 to compute standard errors of the calibrated preference parameters. Additionally, the first approach does not allow to test richer models that could explain behavior through additional elements, 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 (i) uses everyone's full set of decisions to estimate the distribution from where β_i's are drawn; (ii) for each person, uses her decisions to get a conditional distribution from where her β_i is drawn; (iii) uses the mean of the conditional distribution as an estimator of β_i. To simplify the explanation in this section, I will focus on the case where a participant $i \in A$ 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 from picking $d_{im} \in \{Opt1, Opt2\}$ is her base utility function (as defined by equation (1)), plus an idiosyncratic error, $\varepsilon_{d_{im}}$, 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:

$$u(d_{im}) = x_i(d_{im}) + \beta_i \cdot x_j(d_{im}) + \varepsilon_{d_{im}}$$

Where $x_i(d_{im})$ is the payoff i gets when she chooses d_{im} in money allocation decision m.

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:

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

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

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_{s_i}, that has an extreme value distribution with mean zero, and that is independent from $\varepsilon_{d_{im}}$. The expected utility function is:

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

$$u(s_i, s_j) = x_i(s_i, s_j) + \beta_i \cdot x_j(s_i, s_j)$$

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). $x_i(s_i, s_j)$ is the payoff i gets when she chooses s_i and j chooses s_j in the game.

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 other psychological costs, like a distaste for getting what is usually described as the "sucker's payoff" (the payoff i gets when she cooperates and j does not). Ultimately, a positive ψ raises the potential costs of cooperation and, therefore, lowers the level of fear needed to prefer to not cooperate[5]. Appendix B1 reports the process by which I conclude that the expected utility function presented here is the one that best describes behavior. In it, I test alternative models that are more and less general than the one here, test all other possible parameters that would

---

[5]The inclusion of ψ in the model changes the fear threshold for non-cooperation described by equation (2). Now, a person prefers to not cooperate if $\tilde{P}_i(s_j{=}N) \geq \frac{2}{7}\left(\frac{1+5\beta_i}{1+\beta_i+\psi_i}\right)$.

change perceived payoffs to account for any other psychological costs/benefits, and discuss the role of risk aversion. 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:

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

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

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

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

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 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 also 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 based on the money allocation decisions, and 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 elicited individual beliefs, $\tilde{P}_i(s_j{=}N)$.

2.3. Setting of implementation

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, who have lived together for over a hundred years. Throughout much of Jos' history, the relation 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 competing for power began to emphasize religious distinctions, 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 groups being both victims and perpetrators.

After the First Crisis, tensions between the religious groups increased even more. This led to several similar spontaneous outbreaks of inter-religious violence in 2004, 2008 and 2011, each resulting in hundreds of fatalities. The sequence of 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 progressive religious segregation was set in motion, permeating all aspects of life, such as residential areas, schools, jobs, local markets, hospitals and politics.

Figure 2.1 showcases the religious segregation that took place in Jos between 2000 and 2022. The figure shows a diagram for the map of Jos at three key points in time: 2000, just before the First Crisis; 2010, just before the 2011 crisis; and 2022, when the baseline for this study data was collected. In each diagram, each cell represents one of Jos' communities (or

Figure 2.1: Religious Segregation in Jos, 2000-2022
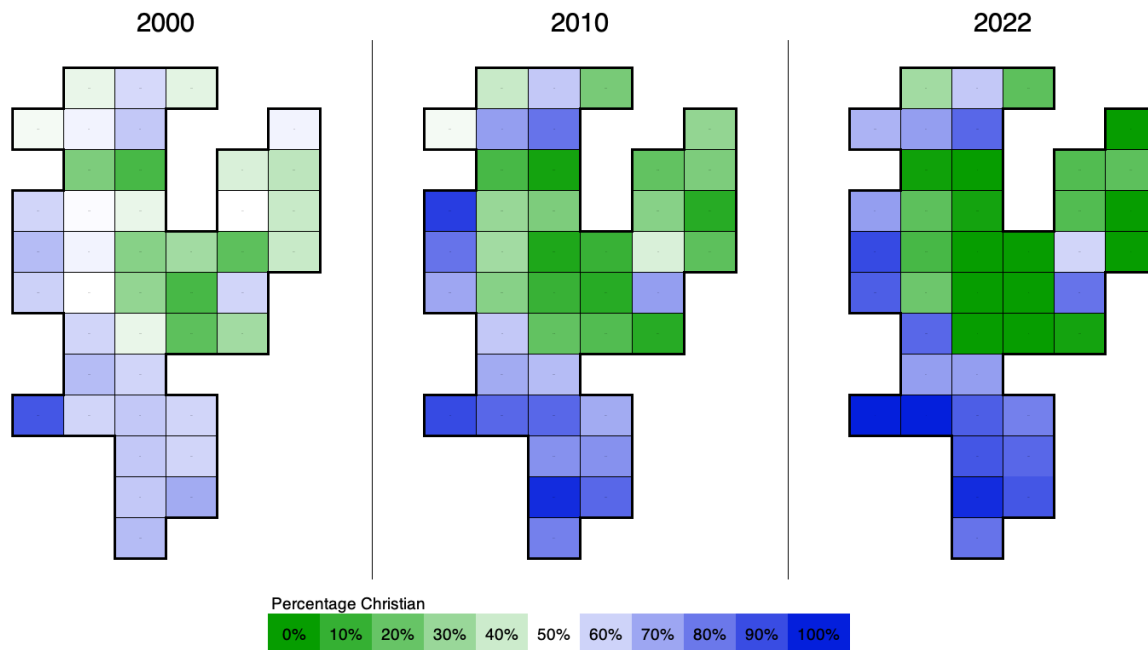
Notes: The figure shows a diagram for the map of Jos at three points in time. In each diagram, each cell represents one of Jos' communities (or neighborhoods) in its approximate geographical location. Green quadrants have a majority Muslim population, with darker greens representing a higher share of Muslims, and blue quadrants have a majority Christian population, with darker blues representing a higher share of Christians. Due to the absence of census data containing this information, the data for this figure was gathered using the following method: 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.

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.

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 religious 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. At the time of this writing, 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. Yet, this equilibrium is constantly under threat and is hard to predict if and when a new crisis may break out.

2.4. Descriptive evidence

The first result is that, in a game where mutual cooperation is an equilibrium and yields 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 leads to a loss of 9.6% of attainable payoffs in intergroup interactions.

Figures 2.2A and 2.2B display 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 2.2A presents a histogram of the social preferences of the participants who decided to not cooperate in the first coordination game of the experiment (the one with $T=-0.2$). 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 2.2A 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, so they must have done so out of fear. It is also worth noting that the rightmost bar indicates that 52% of non-cooperators are, in fact, highly altruistic towards the outgroup.

Figure 2.2B 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. This is, beliefs about the percentage of outgroup members that have a level of hate 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 2.2

A. Social Preferences (for Outgroup) of Non-Cooperators
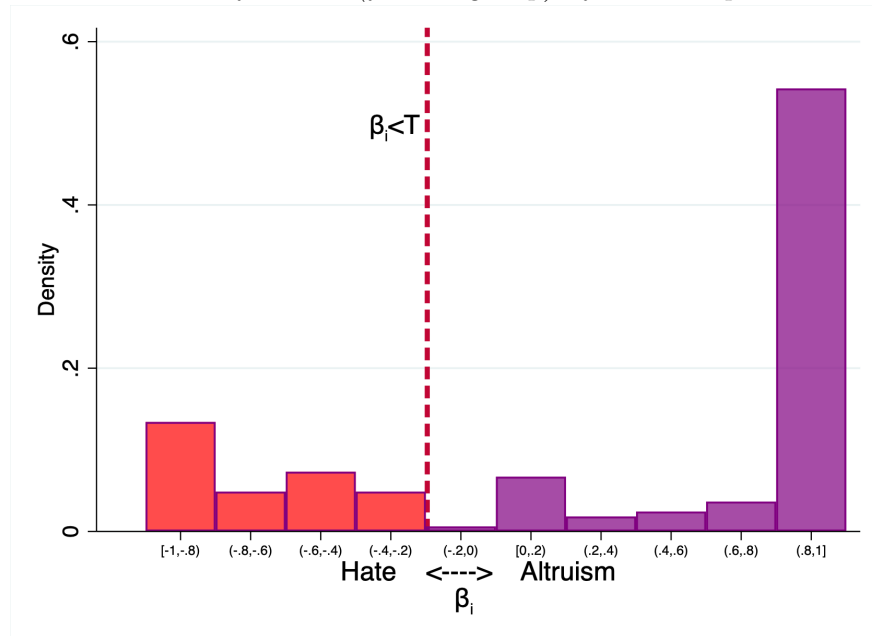
B. Beliefs about the Outgroup, by Game Strategy
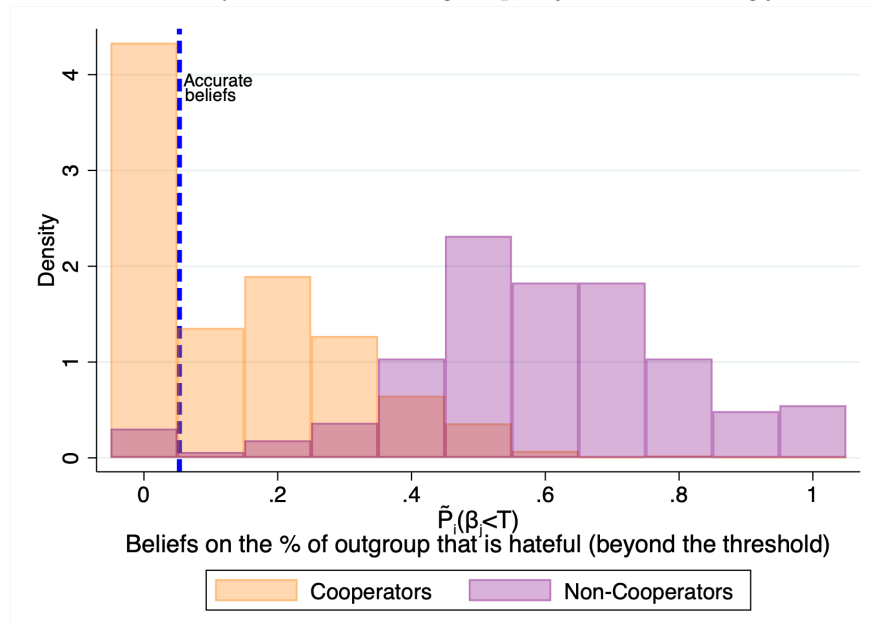
Notes: Figure 2.2A presents the social preferences (β_i) of non-cooperators, estimated following the approach described in Section 3.2. The red line represents the threshold point where a person becomes hateful enough to want to not cooperate out of hate. Figure 2.2B displays the beliefs about probability 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.

Figure 2.3: Preferences, Beliefs and Cooperation

Cooperate   Not Cooperate
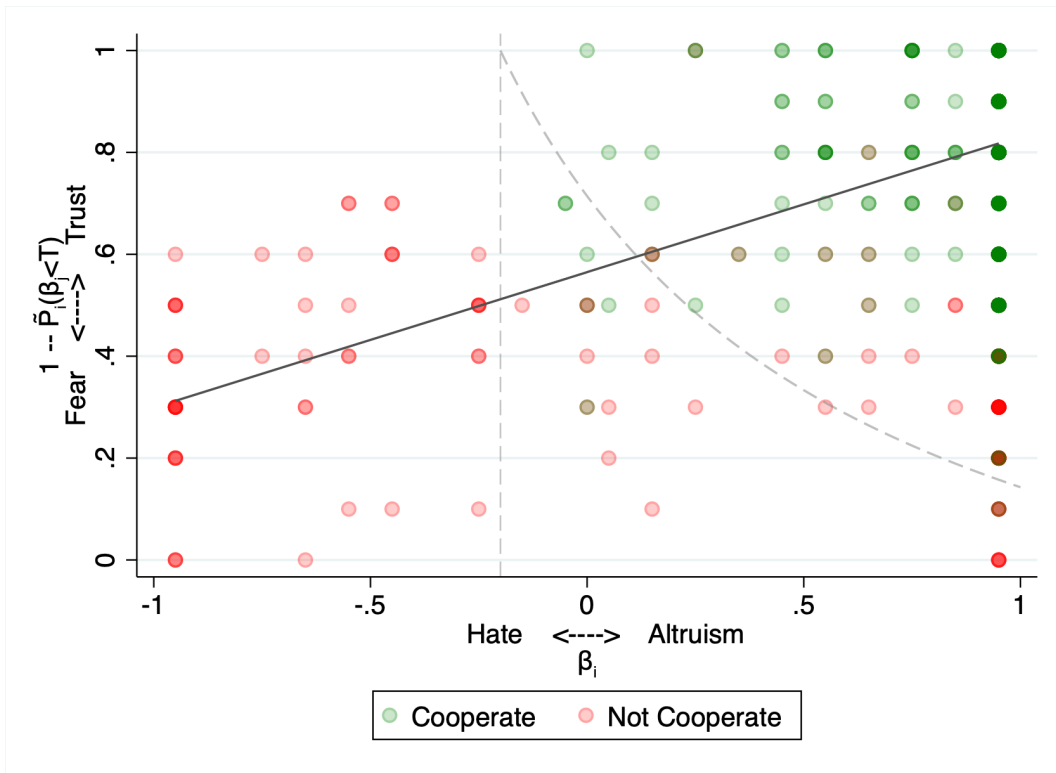
Notes: This figure shows the relationship between preferences, beliefs and cooperation for the entire sample. The x-axis shows participants' social preferences towards the outgroup, β_i, and the y-axis shows participants' beliefs about the outgroup's social preferences towards the ingroup (more specifically, beliefs on the percentage of outgroup members that are not hateful beyond the threshold,$1-\tilde{P}_i(\beta_j<T)$). 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. Tocial preference showcased here were estimated following the approach described in Section 3.2. 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 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 2.3 shows the relationship between preferences, beliefs and cooperation for the entire sample. The x-axis shows participants' social preferences towards the outgroup, β_i, 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, $1-\tilde{P}_i(\beta_j<T)$). 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 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 2.3 shows a few facts worth noticing. First, all of those who fall to the left of the hate threshold do not cooperate. Second, when $\beta_i>T$, the rate of non-cooperation increases as the level of fear 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 display a broader range of beliefs, from complete trust to extreme fear.

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 2.4 reports heterogeneity in social preferences and beliefs.

Figure 2.4A reports the following averages, from left to right: social preferences for the ingroup ($\beta_i|j\in I$), social preferences for the outgroup ($\beta_i|j\in O$), beliefs on the mean social preferences ingroup members have for the ingroup ($\tilde{E}_i[\beta_j|j\in I]$), beliefs on the mean social preferences outgroup members have for the ingroup ($\tilde{E}_i[\beta_j|j\in O]$)—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 for the ingroup than the outgroup. However, the gap between average social preferences for the ingroup vs. the outgroup is much

Figure 2.4: Heterogeneity by Religion

A. Preferences and Beliefs about the Mean

B. Beliefs about the Tail
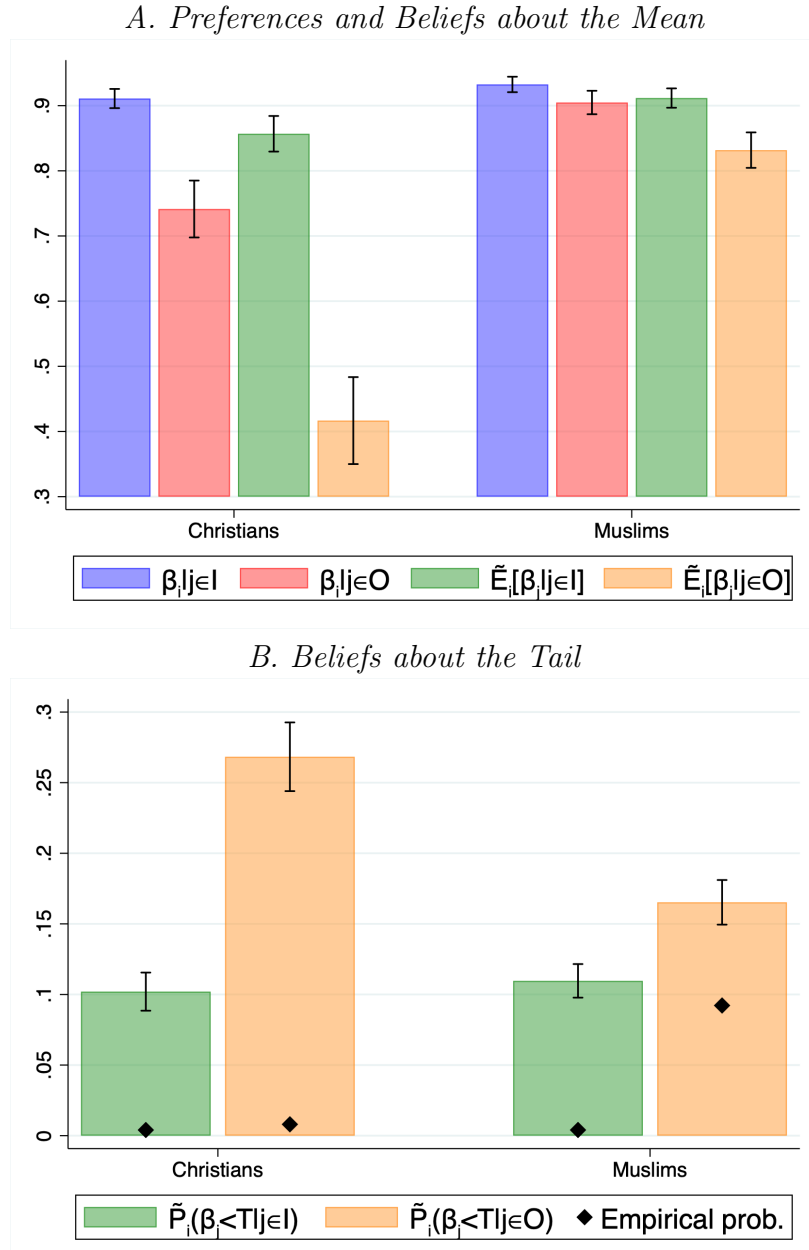
Notes: In both figures, $j \in I$ means that j belongs to the ingroup, and $j \in O$ means that he belongs to the outgroup. Figure 2.4A reports the following averages, from left to right: social preferences for the ingroup $(\beta_i | j \in I)$, social preferences for the outgroup $(\beta_i | j \in O)$, beliefs on the mean social preferences ingroup members have for the ingroup $(\tilde{E}_i[\beta_j | j \in I])$, beliefs on the mean social preferences outgroup members have for the ingroup $(\tilde{E}_i[\beta_j | j \in O])$. Figure 2.4B presents beliefs about the tail of the distribution of social preferences, for Christians and Muslims. $\tilde{P}_i(\beta_j < T | j \in I)$ is the beliefs on the percentage of ingroup members that are hateful beyond the threshold. $\tilde{P}_i(\beta_j < T | j \in O)$ 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.

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 Muslims' social preferences towards them. Instead, if anything, Muslims have somewhat optimistic beliefs about Christians' social preferences towards them.

Figure 2.4B 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, $\tilde{P}_i(\beta_j < T)$. Green bars report beliefs about the ingroup and orange 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 area 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 Christians in Jos that they 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 different willingness to pay.

2.5. Estimated model and conterfactual analysis

Table 2.1 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.

Table 2.1: 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
Likelihood			-3,267

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.

The first thing to note is that all parameters are precisely estimated—all four parameters in Table 2.1 are significant at the 1% level. The estimated μ_β shows that, on average, people are highly altruistic 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 the level of fear needed to want to not cooperate appears to be lower than the one described by equation (2), where the utility function only considers monetary costs. To see this, consider the case of fully altruistic non-cooperators (who are half of all non-cooperators). If $\psi_i=0$ and the only potential costs of cooperating are monetary, a fully altruistic person would want to not cooperate only if she believes that at least 86% of the outgroup will not cooperate ($\tilde{P}_i(s_j=N)>0.86$). However, fully altruistic non-cooperators believe, on average, that 60%

of the outgroup will not cooperate ($\tilde{P}_i(s_j{=}N){=}0.6$). This suggests the fear threshold for non-cooperation is lower than the one described by equation (2), and that, therefore, the potential costs of cooperating go beyond the monetary one. Consequently, a model that includes psychological costs through the ψ parameter will describe better the data.[6]

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 indicates that the core drivers of non-cooperation are captured by my model and measurements.

The quantitative decomposition of motives exercise is done as follows. 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 i's expected utility to i's utility when j cooperates, $W(s_i){=}u(s_i, C)$. 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. It is reassuring to see that the proportions of people not cooperating out of hate and fear estimated by these two approaches are fairly 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 decomposition from both approaches are so similar suggests that the results are not the artifact of a particular specification or

---

[6]Alternatively, this gap could imply that higher-order beliefs are formed and play a role in the cooperation decision. In Appendix B2 I show that this is unlikely to be the case, as preferences and first order beliefs explain almost all the variation in the cooperation decisions.

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 percentage points. 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, i.e., for how many $W(N) > W(C)$. I find that if a policy completely eradicated intergroup hate, the number of people not cooperating would drop by only 5 percentage points. 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 ($\beta_i+0.22$ for those with $\beta_i>0$), 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.

Figure 2.3 can help illustrate these results. In the first counterfactual scenario, all the dots were shifted upwards to match accurate beliefs ($1-\tilde{P}_i(\beta_j<T)$=0.96). As a result, almost all individuals who did not cooperate due to fear (represented by dots on the right of the vertical line and below the curved line) would now have dots above the curved line, which means they no longer have a fearful motive to not cooperate. In the second counterfactual scenario, all the dots to the left of zero were shifted to the right to match β_i=0. In this case, even though the shifted dots now lie to the right of the vertical line and the individuals they represent no longer have a hateful motive to not cooperate, most of them still lie below the curved line, which means these individuals still have a fearful motive to not cooperate.

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 they target (21% vs 96%).

Endline 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.

$$\Delta Cooperate_i = \phi_0 + \phi_1 \Delta Hate_i^{std} + \phi_2 \Delta Fear_i^{std} + \varepsilon_i$$

Here, $\Delta Cooperate_i$ 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. $\Delta Hate_i^{std}$ 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 $\tilde{P}_i(\beta_j<T)$). 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 2.2 reports the results.

Table 2.2: 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 reports a regression where I restricted the sample to subjects for whom only hate changed. The small and insignificant coefficient tells that that changes in hate only were not associated with changes in cooperation. Together with Column 1, these results suggest that reductions in hate can increase cooperation only if accompanied by reductions in fear. On the other hand, Column 3 reports a regression where I restricted the sample to subjects for whom only fear changed. The coefficient 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 coefficient in Column 3 is very similar in magnitude to the one found in Column 1. Hence, the results reported in Table 2.2 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 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.

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

Table 2.3 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 chapter do not merely capture generic negative attitudes but, rather, discern between attitudes underpinned by hatred and those driven by fear.

Table 2.3: 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.

Chapter 3

The Effects of a Cultural Media Intervention on Hate, Fear and Cooperation

After understanding the motives behind social failure, the next step is to study whether policy interventions used in Nigeria are increasing cooperation and why. In this chapter I study cultural media interventions to promote intergroup cooperation. In particular, I focus on radio dramas, an intervention 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 interventions for conflict, radio dramas can be easily implemented in a wide range of contexts. An example of

an alternative policy that has received considerable attention is intergroup contact interventions. But this type of intervention can only be carried out in very particular contexts, where the two communities in conflict live together and tensions are not such that the 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, 2009), indicating the need for further investigation. My goal is to use the experimental protocol to study why these media interventions may or may not be effective at increasing cooperation.

3.1. The radio drama

I partnered with Podbeta, 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, including 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 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 dissertation—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 a 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 it in whichever language 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 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 hatred has a character arc where she heals her resentments, enabling her to form meaningful friendships with members of the other group, fostering reconciliation and unity.

3.2. RCT design

The RCT for the radio show was conducted in the following manner. At first, the surveying company recruited subjects to participate in two lab-in-the-field experiments, which were held two months apart, for a project of the University of California. To minimize experimenter demand effects, enumerators only asked participants if they were interested in taking part in "a different project the surveying company was carrying out" at the conclusion of the initial lab experiment. This second project was for a media production company, and participants were told that it involved listening to a new radio drama that was being released and providing feedback on it. Participants were also informed that they could listen to it at some point before the second lab experiment, which they had already agreed to participate

in. Those who were interested were invited to sign up right away and were told that they would receive additional information through WhatsApp in the following days using the contact number they had given 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 re-measure their preferences, beliefs and cooperation to estimate the effects of the radio drama on each margin.

Half of the participants were assigned to the treatment group and the other half to the control group. These groups were balanced on baseline levels of cooperation, social preferences, beliefs, religion, sex, age and other characteristics. A balance table can be found in Appendix C2. The attrition rate at endline was 5%, which excludes self-selection on this dimension as an issue of the analysis.

3.3. Data collection

The experiments took place in Jos between December 2022 and February 2023. Data collection for each lab-in-the-field experiment (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 lab experiment took around 50 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.

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. Importantly, the final payment did not reveal any information about the participant's or their match's decisions. This was because only some questions, picked at random, got their payoffs implemented, and which question got selected was not revealed. The payments of the quiz lotteries were made directly to the accounts of the winners via phone transfer as soon as the winners were announced.

3.4. Effects of the radio show

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

$$Y_i = \gamma_0 + \gamma_1\,Treated_i + \gamma_2 X_i + \varepsilon_i$$

Where Y_i is an outcome variable; $Treated_i$ 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 3.1 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

Table 3.1: Main Effects of the Radio Show

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 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 effect on the decision to not cooperate in the coordination game.

Table 3.1A 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 have the right sign, towards reducing negative beliefs. Columns 5 and 6 show that the show

had no effects on cooperation either, although again 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 (β_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 the sample to individuals who had margin for improvement in the outcome variable of the respective column. Results are reported in Table 3.1B. 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 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 help illustrate this result. Consider the case of individuals with preferences β_i=−.25, who have hateful preferences just beyond the threshold. Among those with a hateful motive to not cooperate, these are the individuals most likely to be swayed toward cooperation by the intervention. And notice that these participants believe that at least 40% of the outgroup is hateful. The radio show moves 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.

Ultimately, while this result highlights the problems of designing policy interventions without understanding the drivers of non-cooperation, the methodology I offer here shows promise in deepening our understanding of these policy solutions.

Heterogeneous effects by baseline level of hate and fear

I now explore 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 3.2. 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 is 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 in their responses 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 present 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

Table 3.2: Heterogeneous Effects of the Radio Show

	Hate $-\beta_i$	Fear $\tilde{P}_i(\beta_j<T)$	Non-Cooperation $s_i=N$	
	(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.

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 3.3 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

Table 3.3: 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.

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 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

I do not have information on how many people actually listened to the radio show. However, I know 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, but 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 listen

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 Appendix C2. 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.

Conclusion

In this dissertation 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 (even more than hate).

How generalizable are these results to describe intergroup relations? One plausible explanation for unwarranted fears towards the outgroup, which may point at a broader pattern, is that "bad type" 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 (Kahneman, 2011; Enke, 2020)), people will not realize that the information they receive exaggerates the percentage of "bad type" 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 dissertation also suggests that changing beliefs about the outgroup is harder than changing preferences towards the outgroup. In principle, one may think beliefs should be easier to change than preferences. However, there is additional evidence that suggests that in intergroup relations beliefs may be harder to change than preferences. 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, 2020).

However, further research is needed to assess the extent of these findings. Importantly, the experimental protocol in this dissertation 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.

Bibliography

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

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

Adida, C. L., Lo, A., and Platas, M. R. (2018). 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., Baysan, C., Gumren, M., and Kubilay, E. (2021). Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking*. *The Quarterly Journal of Economics*, 136(4):2147–2194.

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

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

Banerjee, A., La Ferrara, E., and Orozco-Olvera, V. H. (2020). The entertaining way to behavioral change: Fighting hiv with mtv.

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

Bauer, M., Cassar, A., Chytilová, J., and Henrich, J. (2014). 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 Miguel, E. (2010). Civil War. *Journal of Economic Literature*, 48(1):3–57.

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

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

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

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

Charness, G. and Chen, Y. (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 Padró i Miquel, G. (2007). Mutual Fear and Civil War.

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

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

Dal Bó, E. and Dal Bó, P. (2014). Do the right thing: The effects of moral suasion on cooperation. *Journal of Public Economics*, 117:28–38.

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

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

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

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

Easterly, W. and Levine, R. (1997). Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics*, pages 1203–1250.

Enke, B. (2020). What you see is all there is. *Quarterly Journal of Economics*, 135(3):1363–1398.

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

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

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

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

Franck, R. and Rainer, I. (2012). 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., Rainer, I., and Trebbi, F. (2015). How is power shared in africa? *Econometrica*, 83(2):465–503.

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

Giuliano, L., Levine, D. I., and Leonard, J. (2009). 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). Ethnic Divisions and Production in Firms. *The Quarterly Journal of Economics*, 129(4):1899–1946.

Hodler, R. and Raschky, P. A. (2014). 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.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Kneeland, T. (2015). Identifying higher-order rationality. *Econometrica*, 83(5):2065–2079.

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

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

Kranton, R., Pease, M., Sanders, S., and Huettel, S. (2020). 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 Goodboy, A. K. (2019). 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). Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration. *American Economic Review*, 111(6):1807–1844.

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

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

Mernyk, J. S., Pink, S. L., Druckman, J. N., and Willer, R. (2022). Correcting inaccurate metaperceptions reduces americans' support for partisan violence. *Proceedings of the National Academy of Sciences*, 119(16):e2116851119.

Michelitch, K. (2015). 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). 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). Does Identity Affect Labor Supply? *American Economic Review*, 113(8):2055–2083.

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

Paluck, E. L. (2009). 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., Green, S. A., and Green, D. P. (2019). The contact hypothesis re-evaluated. *Behavioural Public Policy*, 3(2):129–158. Publisher: Cambridge University Press.

Paluck, E. L., Porat, R., Clark, C. S., and Green, D. P. (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). 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 Warren, S. S. (2018). 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 Rouner, D. (2002). Entertainment-Education and Elaboration Likelihood: Understanding the Processing of Narrative Persuasion. *Communication Theory*, 12(2):173–191.

Tajfel, H. and Turner, J. C. (2004). The social identity theory of intergroup behavior. In *Political Psychology*, pages 276–293. Psychology Press.

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

Tur-Prats, A. and Valencia Caicedo, F. (2020). The long shadow of the spanish civil war.

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

Appendix

Appendix A. Theoretical Model

A2. Derivations of the theoretical model

Derivation of Equation (2)

The following is the derivation of the threshold to not cooperate out of fear represented in equation (2). The condition determines how fearful a person must be, given her level of social preferences, to prefer to not cooperate.

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

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

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.

$$W_i(N) \geq W_i(C)$$

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

$$\tilde{P}_i(s_j{=}N) \cdot (750 + \beta_i 750) \;+\; \tilde{P}_i(s_j{=}C) \cdot (900 + \beta_i 500) \geq$$
$$\tilde{P}_i(s_j{=}N) \cdot (500 + \beta_i 900) \;+\; \tilde{P}_i(s_j{=}C) \cdot (1000 + \beta_i 1000)$$

Replacing $\tilde{P}_i(s_j{=}C) = 1 - \tilde{P}_i(s_j{=}N)$ we get

$$350 \cdot \tilde{P}_i(s_j{=}N) + 350 \cdot \beta_i \cdot \tilde{P}_i(s_j{=}N) \geq 100 + 500 \cdot \beta_i$$
$$\tilde{P}_i(s_j{=}N) \cdot 350 \cdot (1 + 5\beta_i) \geq 100 \cdot (1 + 5\beta_i)$$
$$\tilde{P}_i(s_j{=}N) \geq \frac{2}{7}\left(\frac{1 + 5\beta_i}{1 + \beta_i}\right)$$

Appendix B. Empirical Model

B1. Alternative Empirical Models

For all the following empirical models I test, the utility function in the money allocation decisions is:

$$u(d_{im}) = x_i(d_{im}) + \beta_i \cdot x_j(d_{im}) + \varepsilon_{d_{im}}$$

Where $d_{im} \in \{Opt1, Opt2\}$, $x_i(d_{im})$ is the payoff i gets when she chooses d_{im} in money allocation decision m, and $\varepsilon_{d_{im}}$ is an idiosyncratic error that has an extreme value distribution with mean zero. 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.

In what follows I test different empirical models in which I vary the expected utility functions that would be used in the coordination game.

Most restricted expected utility function

The first model I test has an expected utility function as defined in theoretical model of Section 3.1. The expected utility function is the following.

$$W(s_i) = \tilde{P}_i(s_j{=}C){\cdot}u(s_i, s_j{=}C) \ + \ \tilde{P}_i(s_j{=}N){\cdot}u(s_i, s_j{=}N) + \varepsilon_{s_i}$$
$$u(s_i, s_j) = x_i(s_i, s_j) + \beta_i \cdot x_j(s_i, s_j)$$

Where $\tilde{P}_i(s_j{=}s)$ is i's subjective beliefs on $P(s_j{=}s)$, given that $j{\in}B$, $x_i(s_i, s_j)$ is the payoff i gets when she chooses s_i and j chooses s_j in the game, and ε_{s_i} is an idiosyncratic error that has an extreme value distribution with mean zero. 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 parameter is β_i.

I use a random coefficient model to estimate the parameters of the distribution of β_i (mean μ_β, and variance σ_β), as explained in Section 4. The results are reported in Table B1.1.

Table B1.1: Most Restricted Expected Utility

	Coefficient	Stand. Err.	
μ_β	0.951	0.064	***
σ_β	0.447	0.062	***
Observations		9,006	
Clusters		997	
Likelihood		-3,497	

Notes: This table reports the results of the simulated maximum likelihood estimation of a random coefficients model. 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. Standard errors are clustered at the individual level.

The in-sample fit performance of this model is 90% for the full sample and 43% for non-cooperators. In other words, the parameters estimated correctly predict the decisions taken by 90% of participants in the experiment and 43% of participants that decided to not-cooperate in the experiment. The reason why this model does a poor job at explaining non-cooperative behavior is that half of the people who did not cooperate are fully altruistic, as shown in Figure 2. If $\psi_i=0$, 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. To account for this and increase the model fit in non-cooperators, I test more flexible models.

Risk aversion

One first approach to relax the initial model assumptions would be to have an expected utility function that allows for 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 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). Because of this, the first variation of the model I test is one with loss aversion.

Loss aversion

In search for a better model fit with the data, I relax the functional assumption and consider a model that introduces loss aversion. In the game, given its strategic nature, there is uncertainty in the outcome, and each strategy represents facing a different lottery. Choosing to not cooperate minimizes the losses of a player by guarantees a minimum payoff of 750. In this case, i can be positively surprised and get 900, if j decides to cooperate. Alternatively, choosing to cooperate aims to maximize earning, allowing the player to reach the maximum potential payoff of 1000. In this case, i can be negatively surprised and get 500 if j decides to not cooperate. According to prospect theory, a negative surprise represent a greater change in utility than a positive surprise in the same amount. In line with this, there is a psychological cost of cooperating when the other person does not cooperate that goes beyond the monetary payoff. To account for this, I test the following utility function with loss aversion.

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

$$u(s_i, s_j) = x_i(s_i, s_j) + \beta_i \cdot x_j(s_i, s_j)$$

Where $\tilde{P}_i(s_j{=}s)$ is i's subjective beliefs on $P(s_j{=}s)$, given that $j{\in}B$, $x_i(s_i, s_j)$ is the payoff i gets when she chooses s_i and j chooses s_j in the game, and ε_{s_i} is an idiosyncratic error that has an extreme value distribution with mean zero.

ψ is a parameter of loss aversion, where the reference point is the payoff from *(Cooperate, Cooperate)*. In addition, the psychological cost ψ captures can go beyond loss aversion. 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).

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. Notice also that the first model is nested in this one.

I use a random coefficient model to estimate the parameters of the distribution of β_i (mean μ_β, and variance σ_β) and the parameter ψ, as explained in Section 4. The results are reported in Table B1.2.

The first thing to note is that the estimated parameter ψ is significant, which indicates that is seems to matter in the decision problem. The in-sample fit performance of this model is 96% for the full sample and 91% for non-cooperators. In other words, the parameters estimated correctly predict the decisions taken by 96% of participants in the experiment and 91% of participants that decided to not-cooperate in the experiment. The model fit for non cooperators increases substantially when compared to the previous model. This reveals that loss aversion (or the psychological cost of getting the "sucker's payoff" does seem to play a

Table B1.2: Loss Aversion

	Coefficient	Stand. Err.	
μ_β	0.936	0.071	***
σ_β	0.419	0.061	***
ψ	539.4	108.2	***
Observations		9,006	
Clusters		997	
Likelihood		-3,299	

Notes: This table reports the results of the simulated maximum likelihood estimation of a random coefficients model. 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. ψ is a parameter of loss aversion. Standard errors are clustered at the individual level.

role in the decision to cooperate. In the end, this model is empirically superior to the one without loss aversion.

Multiple payoff shifters

Ultimately, the parameter ψ from the previous model is a payoff shifter that changes the way in which i perceives the payoffs from each strategy in the case where j does not cooperate. This begs the question if a symmetric payoff shifter for the case where j cooperates would also increase the explanatory power of the model. Such a payoff shifter could represent a psychological benefit various sorts. For one, it could represent a taste for mutual cooperation, if this weight is positive when i decides cooperates. Alternatively, it could represent a taste for dominance (i.e., getting a higher payoff than the other player) if this weight is negative when i decides coopertes. In the following model I test if including an additional payoff shifter increase the explanatory power of the model. The expected utility function is the following:

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

$$u(s_i,s_j) = x_i(s_i,s_j) + \beta_i \cdot x_j(s_i,s_j)$$

Where $\tilde{P}_i(s_j{=}s)$ is i's subjective beliefs on $P(s_j{=}s)$, given that $j{\in}B$, $x_i(s_i,s_j)$ is the payoff i gets when she chooses s_i and j chooses s_j in the game, and ε_{s_i} is an idiosyncratic error that has an extreme value distribution with mean zero.

Additionally, ψ is a payoff shifter for the case when j does not cooperate, that can capture loss aversion or other psychological costs. And γ is a payoff shifter for the case when j cooperates that can capture psychological benefits like a taste for cooperate or dominance.

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, ψ and γ. Notice also that the first two models are nested in this one.

I use a random coefficient model to estimate the parameters of the distribution of β_i (mean μ_β, and variance σ_β) and the parameters ψ and γ, as explained in Section 4. The results are reported in Table B1.3.

B1.3: Multiple Payoff Shifters

	Coefficient	Stand. Err.	
μ_β	0.938	0.069	***
σ_β	0.420	0.059	***
ψ	565.4	177.2	***
γ	25.8	127.8	
Observations		9,006	
Clusters		997	
Likelihood		-3,299	

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. ψ is a parameter of loss aversion. γ is a payoff shifter. Standard errors are clustered at the individual level.

The first thing to note is that the parameter γ is not significant, and the likelihood this model reaches is not greater than the previous model that had no γ. This indicates that there does not seem to be a psychological benefit that shift the perceived payoffs in the case when j cooperates.

The fact can shifter ψ is significant and γ is not can be understood by considering Figure 3. The payoff shifter γ would move the cooperation threshold represented by the vertical line. Instead, the payoff shifted ψ would move the cooperation threshold represented by the curved line. As one can see graphically, moving the vertical line would not improve the fit, because to the left of it there are only non-cooperators (as the theory predicts). Instead, moving the curved line with the payoff shifter ψ would improve the sample fit because above

this curved line there are some non-cooperators that the base expected utility function is not explaining.

In conclusion, only one psychological cost/benefit that seems to be playing a role in the decision problem. In this sense, the previous model, with ψ and no γ, is empirically superior to this one.

Loss aversion at the individual level

Given that the estimation occurs in the environment of random coefficients, one can easily increase the flexibility of the model to get a better fit by taking ψ to be a random coefficient too, not just β_i. In this sense, every person has a different level of loss aversion ψ_i that comes from a distribution with mean μ_ψ and variance σ_ψ^2. The expected utility function is the following.

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

$$u(s_i, s_j) = x_i(s_i, s_j) + \beta_i \cdot x_j(s_i, s_j)$$

Where $\tilde{P}_i(s_j{=}s)$ is i's subjective beliefs on $P(s_j{=}s)$, given that $j{\in}B$; $x_i(s_i, s_j)$ is the payoff i gets when she chooses s_i and j chooses s_j in the game; ψ is an individual-level parameter of loss aversion; and ε_{s_i} is an idiosyncratic error that has an extreme value distribution with mean zero.

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.

I use a random coefficient model to estimate the parameters of the distributions of β_i (mean μ_β, and variance σ_β) and ψ_i (mean μ_ψ, and variance σ_ψ), as explained in Section 4. The results are reported in Table B1.4.

The first thing to note is that the estimated parameters of ψ's distribution are significant, which indicates that there seems to be some individual level variation in loss aversion. The in-sample fit performance of this model is 99% for the full sample and 95% for non-cooperators. In other words, the parameters estimated correctly predict the decisions taken by 99% of participants in the experiment and 95% of participants that decided to not-cooperate in the experiment. Compared to the other model with loss aversion, this model presents a better sample fit and reaches a higher likelihood. Even if the gains aren't very big, the extra flexibility this model presents does seem to help better fit the data. In this sense, this model is empirically superior to the one individual-level loss aversion.

This is then the preferred model over all, as it seems to be the one to best represent the behavior observed in the lab. Because this, is the the model selected for the paper and the one presented in Section 4.

Table B1.4: Individual-Level Loss Aversion

	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	
Likelihood		-3,267	

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.

B2. Higher Order Beliefs

As mentioned earlier, the estimated model with the base utility function only predicts correctly the choices of 43% of non-cooperators. Although loss aversion seems to account quite well for what that model is missing, an alternative explanation is that in the model I should to account for higher order beliefs in order to get a better sample fit. In what follows I show evidence of why it is unlikely higher order beliefs are playing a role in the lab decisions. In addition, I estimate the upper bound of how much could higher order beliefs matter if I relaxed this assumption.

**Evidence that participants do not form higher-order beliefs in the lab game**

The first piece of evidence that supports that participants do not form higher-order beliefs in the lab game is the fieldwork done during the pilot. When talking to participants to understand their reasoning behind their decision in the game, no explanation that resembled higher-order beliefs ever came up. If the participant gave a justification on why they thought their match would not cooperate, it always had to do with the belief that their match wanted to lower their payoff. When asked specifically about higher order beliefs in the game, participant struggle a lot to understand (or did not understand at all) how beliefs on beliefs should be affecting their decision and that of their partner. It is worth remarking that

this was the first time the participants faced a game from game theory, so understanding the basics of the game was already demanding enough.

Importantly, this anecdotal evidence from the fieldwork falls in line with the ample evidence in the experimental literature that states that, on average, people do not form higher order beliefs when playing games (Rubinstein, 1989; Kneeland, 2015). And what is more, they are less likely to form higher order beliefs if their are playing a game for the first time (as was the case in my setting).

Beyond this, the data in of the experiment does not seem to suggest that higher order beliefs are part of participants decisions. Assuming higher order beliefs are not perfectly correlated with first order beliefs, one should expect that people with the same preferences and first order beliefs could behave differently because they could have different higher order beliefs. However, the data shows that preferences and first order beliefs seem to explain almost all the variation in behavior. Figure B2.1 shows evidence on this.

Figure B2.1: Variance in Cooperation Decisions for each Preferences-First Order Beliefs
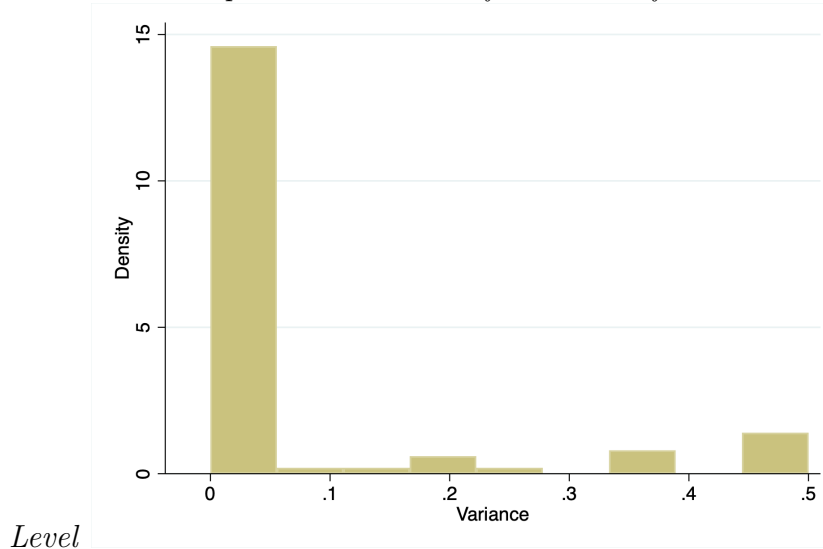
Level

The figure reports how much variance there is in cooperation decisions within a each preference-first order beliefs level. I find that there is very little variance in behavior when controlling for level of preferences and first order beliefs. 77% has zero variance of the levels show zero variance in decisions.

This is evidence that the model lack of fit with the data is probably not coming from misspecified beliefs, but instead from misspecified thresholds of cooperation. As shown in the previous section, the thresholds of cooperation are changed with payoff shifters. Indeed, as shown in the previous section, changing the thresholds of cooperations (in particular, that

of fear) through payoff shifters proves to be rather effective at increasing the sample fit of the model.

Upper bound of higher-order beliefs

As argued before, it is unlikely that higher order beliefs are playing a role in the decisions in the game. Regardless, one can explore how big would their role be if we assumed that all the lack of fit from the model all comes from only higher order beliefs. This is a very strong assumption, as it assumes that participants have no risk aversion, loss aversion or psychological costs/benefits, so higher order beliefs is the only thing that can explain the wrong predictions of the model with only first order beliefs. But this cans till be an illustrative exercise.

Under this assumption, I calculate for each person how much more fear would they need to have in order to act the way the acted. Specifically, I find how many percentage points of fear more do each person need to explain their behavior. With this information, I can calculate what percentage of the fear needed to act in the way a person acted should be coming from higher-order beliefs. I find that level of fear needed for observed behavior that is unaccounted for by first order would be 3% for the full sample, 19% for all non-cooperators, and 27% for non-cooperators without a hateful motives to not cooperate, only a fearful one. This puts an upper bound on the role higher order beliefs could have. However, for the reasons explained above, it is unlikely that this is the case.

Appendix C. Results

C1. Social Desirability Bias in the Measure of Hate Measure

One potential concern is that the elicited hate is affected by social desirability bias, and that this, in turn, biased the previous analysis. Despite the choices made on the experimental design to reduce demand effects, one might be concerned that participants realize the experiment is about religious division, and, considering this, might disingenuously express more positive attitudes towards the outgroup to present themselves in a good light to the surveyors. To check how robust the previous analysis is to social desirability bias, I do the following. In the baseline survey, I include 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 social desirability scores (SDS)

to correct the estimated social preferences in the following way. Let the biased estimated social preferences be:

$$\beta_{i,\text{biased}} = \beta_{i,\text{true}} + \phi \text{SDS}_i$$

I estimate ϕ by regressing $\beta_{i,\text{biased}}$ on SDS_i. Then, to get $\beta_{i,\text{true}}$, I compute $\beta_{i,\text{true}} = \beta_{i,\text{biased}} - \phi \text{SDS}_i$ for each person. I then input these new social preferences into the model and see how the analysis changes.

SDS is a measure from 0 to 10, where 10 represents having reported that one has all 10 too-good-to-be-true traits. The estimated ϕ is XX, which means that people one unit of SDS means that the β_i is overestimated by 0.2 units.

Results are mostly unchanged. Non-cooperation increases by XXX% and the motives for non-cooperation are now XXX% out of hate and XXX% out of fear. The reason why results are mostly unchanged is intuitive: people with SDS report very high β_i. That means that even when adjusting their parameter, it is not enough for them to want to not cooperate. And for those non-cooperators out of fear, same things, adjusting is not enough for them to want to not cooperate out of hate. Their reason is maintained, they don't cooperate out of fear.

C2. Additional Results of the RCT

Table C2.1: Balance on Observables

	(1) No-Coop	(2) Soc.Pref.	(3) Beliefs	(4) Religion	(5) Gender	(6) Age	(7) SDS
Treated	0.014	-0.004	0.004	-0.005	0.003	0.022	0.117
	(0.024)	(0.025)	(0.015)	(0.033)	(0.032)	(0.733)	(0.131)
Observations	947	947	947	947	947	947	947

Table C2.2: Average Treatment Effect on the Treated

	(1)	(2)	(3)
	Soc.Pref.	Beliefs	No-Coop.
Treated	0.080**	-0.038	-0.043
	(0.037)	(0.032)	(0.043)
Controls	Y	Y	Y
Mean Dep.Var.	.823	.218	.169
Observations	947	947	947

Table C2.3: Radio Show Effects on Policy Attitudes

	(1)	(2)	(3)	(4)
	Settlements	Schools	Hospitals	Markets
Treated	0.013	-0.037	-0.009	0.032
	(0.049)	(0.046)	(0.028)	(0.029)
Controls	Y	Y	Y	Y
Mean Dep.Var.	2.815	3.218	3.655	3.634
Observations	942	941	946	946

Appendix D. Lab Experiment

D1. Screenshots and Pictures

Green and Blue groups in the Lab Game

BLUE GROUP

Aaron	Joshua
Mercy	Paul
Titus	Mary
Ruth	Isaac
Blessing	Grace

GREEN GROUP

Abubakar	Umar
Zainab	Ahmadu
Tijani	Maimuna
Aisha	Shehu
Hauwa	Zuwaira

Examples of Money Allocation Decisions

Eliciting Beliefs About the Outgroup, Example
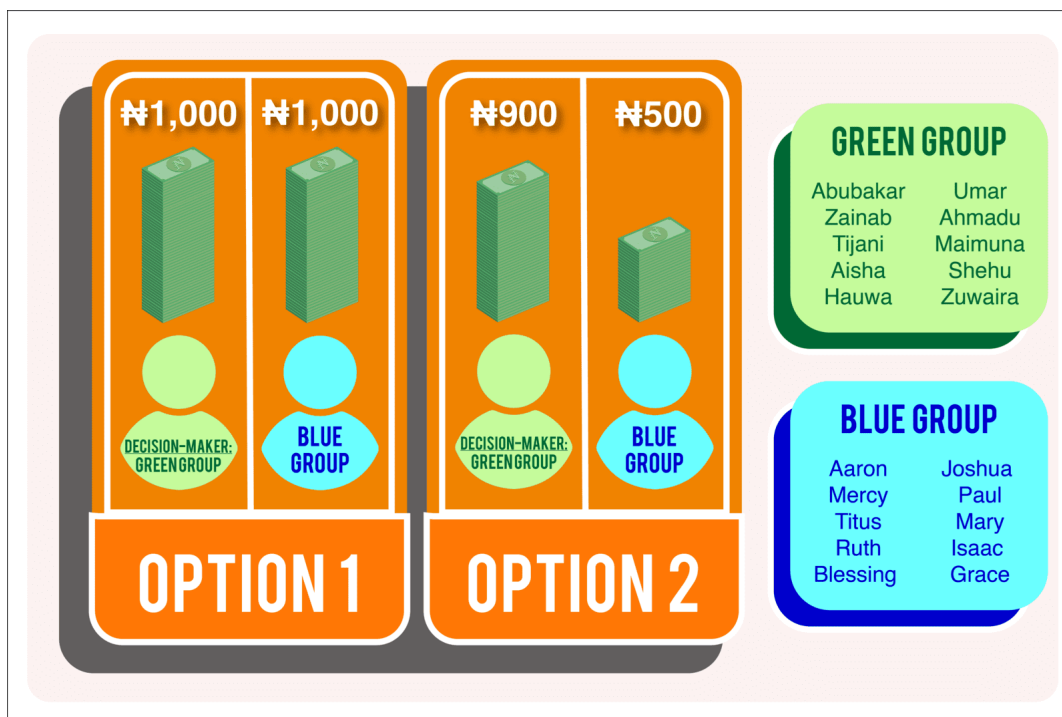
How many people from the Green Group do you think picked Option 2?   (1—10)

Coordination Game Example

GREEN GROUP

Abubakar	Umar
Zainab	Ahmadu
Tijani	Maimuna
Aisha	Shehu
Hauwa	Zuwaira

YOUR choice	Match's choice	YOUR earnings	Match's earnings
Cooperate	Cooperate	₦1,000	₦1,000
Cooperate	Not Cooperate	₦500	₦900
Not Cooperate	Cooperate	₦900	₦500
Not Cooperate	Not Cooperate	₦750	₦750

Lab in the Field

Jos, Nigeria

D2. Experimental Protocol

The following is a sketch of experimental protocol that is best for illustration purposes. The protocol used in the field had multiple intermediate steps to guarantee participants got a good understanding of the games.

Introduction

The survey you are going to take today is anonymous. We won't record your name at any point, so no one will be able to link your choices to you. We will never share individual answers, only aggregate statistics, so you should feel free to answer whatever you want.

The results from this survey will be analyzed and shared only in the United States, never in Nigeria. No Nigerian organization, agency or researcher is involved.

Do you have any questions or concerns about the anonymity or confidentiality of your answers?

In this survey, there are no wrong or right answers, there is no objective to achieve. Different people prefer different things, and that is absolutely fine.

Today we will go through several activities that will together take about 50 minutes to complete. For today's survey you will receive between ₦700 and ₦1700, depending on the decision you and other participants in this study make. It is in your best interest to pay close attention to each question.

Your earnings will be paid to you in cash right after we finish the survey.

Module 1: Demographics

Religion, Gender, Age, Education, Migration, etc.

Module 2: Social Preferences (Money Allocation Decisions) with the First Group

Now we will proceed with the activities.

One month ago we gathered a group of people from different parts of Jos, and ages from 18 to 60. We chose twenty of them and divided them into two groups, the Blue Group and the Green Group. They participated in multiple activities, similar to the ones you are about to go through.

In these activities, you will make decisions that can make you, and them, earn extra money. Participants from the Blue Group and the Green Group that receive extra money from your decisions will never know your identity, or what your decisions were. They will only find out what their total extra earnings was, which will be affected by multiple things in addition to your decisions. Their extra earnings will be paid to them in the following days.

From all the decisions you make that could earn you extra money, we will randomly pick only one and implement its payments. So you should consider each decision on its own, and not think of them as accumulating.

If you don't have any question we will now procede with the activities.

[The first group the participant plays with is picked at random. In this example, the participant will play first with the Blue group]

In this first activity you will be randomly matched with a member from the Blue Group. To keep his/her anonymity, you will exactly know who he/she is. All you will know is that he/she is a person from the Blue Group, which consists of the following ten people:

<div align="center">

BLUE GROUP

ChristianName1	ChristianName2	ChristianName3
ChristianName4	ChristianName5	ChristianName6
ChristianName7	ChristianName8	ChristianName9
	ChristianName10	

</div>

In this task you will make a series of decisions that will make both of you earn money. However, how much money you and your match get depends only on a decision you will make. We want to understand how much of your earnings you are willing to give up in order to increase or decrease in ₦500 the earnings of your match.

Remember, we will implement the payments of at most one of these decisions (picked randomly), so you should not think of them as accumulating or compensating one another. The best thing for you to do is to consider each of these decisions on its own.

[Participant go through 7 or 8 of the following decisions, following the algorithm detailed in this Appendix]

(Altruism, p=0)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦0
Option 2 — You get ₦1,000, and your match from the Blue Group gets ₦500

(Altruism, p=50)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦0
Option 2 — You get ₦950, and your match from the Blue Group gets ₦500

(Altruism, p=100)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦0
Option 2 — You get ₦900, and your match from the Blue Group gets ₦500

(Altruism, p=150)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦0
Option 2 — You get ₦850, and your match from the Blue Group gets ₦500

(Altruism, p=200)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦0
Option 2 — You get ₦800, and your match from the Blue Group gets ₦500

(Altruism, p=250)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦0
Option 2 — You get ₦750, and your match from the Blue Group gets ₦500

(Altruism, p=300)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦0
Option 2 — You get ₦700, and your match from the Blue Group gets ₦500

(Altruism, p=350)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦0
Option 2 — You get ₦650, and participant from the Blue Group gets ₦500

(Altruism, p=400)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦0
Option 2 — You get ₦600, and your match from the Blue Group gets ₦500

(Altruism, p=450)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦0
Option 2 — You get ₦550, and your match from the Blue Group gets ₦500

(Hate, p=0)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦1,000

Option 2 — You get ₦1,000, and your match from the Blue Group gets ₦500

(Hate, p=50)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦1,000
Option 2 — You get ₦950, and your match from the Blue Group gets ₦500

(Hate, p=100)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦1,000
Option 2 — You get ₦900, and your match from the Blue Group gets ₦500

(Hate, p=150)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦1,000
Option 2 — You get ₦850, and your match from the Blue Group gets ₦500

(Hate, p=200)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦1,000
Option 2 — You get ₦800, and your match from the Blue Group gets ₦500

(Hate, p=250)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦1,000
Option 2 — You get ₦750, and your match from the Blue Group gets ₦500

(Hate, p=300)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦1,000
Option 2 — You get ₦700, and your match from the Blue Group gets ₦500

(Hate, p=350)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦1,000
Option 2 — You get ₦650, and your match from the Blue Group gets ₦500

(Hate, p=400)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦1,000
Option 2 — You get ₦600, and your match from the Blue Group gets ₦500

(Hate, p=450)
Option 1 — You get ₦1,000, and your match from the Blue Group gets ₦1,000
Option 2 — You get ₦550, and your match from the Blue Group gets ₦500

Module 3: Social Preferences (Money Allocation Decisions) with the Second Group

Now you are going to be matched at random with a different person, this time a member from the Green Group. Again, to keep his/her anonymity, you won't exactly know who he/she is. All we will tell you is that he/she is a member of the Green Group, which consists of the following ten people:

GREEN GROUP

MuslimName1	MuslimName2	MuslimName3
MuslimName4	MuslimName5	MuslimName6
MuslimName7	MuslimName8	MuslimName9
	MuslimName10	

As before, your task is to pick if you and your match from the Green Group get the money in Option 1 or Option 2. We want to understand how much of your earnings you are willing to give up in order to increase or decrease in ₦500 the earnings of your match.

Remember, we will implement the payments of at most one of these decisions (picked randomly), so you should not think of them as accumulating or compensating one another. The best thing for you to do is to consider each of these decisions on its own.

[Participants then go again through the money allocation decisions, now with their match from the group they haven't played with (in this example, the Green Group)]

Module 4: Beliefs on the Social Preferences of the Outgroup

[In this example, the outgroup is the Green Group]

This next task is about guessing what other participants in this survey did. If you guess correctly in all questions, we will add ₦500 to your earnings.

A month ago, people from the Green Group and the Blue Group went through the same activity you just went through. To each member of the Green Group we told them that we had matched them at random with someone from the Blue Group, and showed them the list of names of the Blue Group. Then, they proceeded to make the same series of decisions you just made: picking between the money in Option 1 or Option 2.

In this task, you have to guess what a member of the Green Group picked when playing with his/her match from the Blue Group.

What do you think the member of the Green Group picked when matched with a member of the Blue Group?

[Participants made guesses for at most eight of the following money allocation decision, following the same algorithm as before]

(Altruism, p=0)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦0
Option 2 — Participant from the Green Group decided for him to get ₦1000, and his match from the Blue Group to get ₦500

(Altruism, p=50)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦0
Option 2 — Participant from the Green Group decided for him to get ₦950, and his match from the Blue Group to get ₦500

(Altruism, p=100)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦0
Option 2 — Participant from the Green Group decided for him to get ₦900, and his match from the Blue Group to get ₦500

(Altruism, p=150)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦0
Option 2 — Participant from the Green Group decided for him to get ₦850, and his match from the Blue Group to get ₦500

(Altruism, p=200)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦0
Option 2 — Participant from the Green Group decided for him to get ₦800, and his match from the Blue Group to get ₦500

(Altruism, p=250)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦0
Option 2 — Participant from the Green Group decided for him to get ₦750, and his match

from the Blue Group to get ₦500

(Altruism, p=300)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦0
Option 2 — Participant from the Green Group decided for him to get ₦700, and his match from the Blue Group to get ₦500

(Altruism, p=350)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦0
Option 2 — Participant from the Green Group decided for him to get ₦650, and his match from the Blue Group to get ₦500

(Altruism, p=400)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦0
Option 2 — Participant from the Green Group decided for him to get ₦600, and his match from the Blue Group to get ₦500

(Altruism, p=450)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦0
Option 2 — Participant from the Green Group decided for him to get ₦550, and his match from the Blue Group to get ₦500

(Hate, p=0)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000
Option 2 — Participant from the Green Group decided for him to get ₦1000, and his match from the Blue Group to get ₦500

(Hate, p=50)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000
Option 2 — Participant from the Green Group decided for him to get ₦950, and his match from the Blue Group to get ₦500

(Hate, p=100)
Option 1 — Participant from Group B decided for him/her to get ₦1,000, and his/her match

from the Blue Group to get ₦1,000

Option 2 — Participant from Group B decided for him to get ₦900, and his match from the Blue Group to get ₦500

(Hate, p=150)

Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000

Option 2 — Participant from the Green Group decided for him to get ₦850, and his match from the Blue Group to get ₦500

(Hate, p=200)

Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000

Option 2 — Participant from the Green Group decided for him to get ₦800, and his match from the Blue Group to get ₦500

(Hate, p=250)

Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000

Option 2 — Participant from the Green Group decided for him to get ₦750, and his match from the Blue Group to get ₦500

(Hate, p=300)

Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000

Option 2 — Participant from the Green Group decided for him to get ₦700, and his match from the Blue Group to get ₦500

(Hate, p=350)

Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000

Option 2 — Participant from the Green Group decided for him to get ₦650, and his match from the Blue Group to get ₦500

(Hate, p=400)

Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000

Option 2 — Participant from the Green Group decided for him to get ₦600, and his match from the Blue Group to get ₦500

(Hate, p=450)
Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000
Option 2 — Participant from the Green Group decided for him to get ₦550, and his match from the Blue Group to get ₦500

[The next questions asked about their beliefs on the tail of the distribution of social preferences and the dispersion]

Now I will ask you to make a couple more guesses:

(Tail, p=100)
In the following decision, how many people from the Green Group do you think picked Option 2 when deciding how much money they and their match from the Blue Group would get? [0—10]

Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000
Option 2 — Participant from the Green Group decided for him to get ₦900, and his match from the Blue Group to get ₦500

(Tail, p=300)
In the following decision, how many people from the Green Group do you think picked Option 2 when deciding how much money they and their match from the Blue Group would get? [0—10]

Option 1 — Participant from the Green Group decided for him/her to get ₦1,000, and his/her match from the Blue Group to get ₦1,000
Option 2 — Participant from the Green Group decided for him to get ₦700, and his match from the Blue Group to get ₦500

(Dispersion 1)
How sure do you feel that your guesses in the past module are correct? 1—5
1 – Not sure at all; 5 – Absolutely sure.

(Dispersion 2)
How similarly do you think the choices of people from the Green Group were in this type of decisions? 1—5

1 – They all answered in the same way; 5 – They answered in all sorts of ways.

Module 5: Beliefs on the Social Preferences of the Ingroup

As in the previous task, this next task is about guessing what other participants in this survey did. If you guess correctly in all questions, we will add ₦500 to your earnings.

At some point in last month session, we told the members from the Blue Group that we had matched them at random with someone from inside the Blue Group, and showed them the list of names. Then, they proceeded to make the same series of decisions you just made: picking between the money in Option 1 or Option 2.

In this task, you have to guess what a member of the Blue Group picked when playing with his/her match from inside the Blue Group.
[Here they proceed to guess the money allocation decisions of an ingroup member]

Now I will ask you to make a couple more guesses:

(Tail, p=100)
In the following decision, how many people from the Blue Group do you think picked Option 2 when deciding how much money they and their match from inside the Blue Group would get? [0—10]

Option 1 — Participant from the Blue Group decided for him/her to get ₦1,000, and his/her match from inside the Blue Group to get ₦1,000
Option 2 — Participant from Blue Group decided for him to get ₦900, and his match from the inside the Blue Group to get ₦500

(Tail, p=300)
In the following decision, how many people from the Blue Group do you think picked Option 2 when deciding how much money they and their match from inside the Blue Group would get? [0—10]

Option 1 — Participant from the Blue Group decided for him/her to get ₦1,000, and his/her match from inside the Blue Group to get ₦1,000
Option 2 — Participant from Group B decided for him to get ₦700, and his match from inside the Blue Group to get ₦500

(Dispersion 1)

How sure do you feel that your guesses in the past module are correct? 1—5

1 – Not sure at all; 5 – Absolutely sure.

(Dispersion 2)

How similarly do you think the choices of people from the Blue Group were in this type of decisions? 1—5

1 – They all answered in the same way; 5 – They answered in all sorts of ways.

Module 6: Coordination Games with the Ingroup and Outgroup

In this example, the Blue Group is the ingroup and the Green Group is the outgroup]

In the next activity you and your match have to make a decision, and ¡b¿the monetary outcome depends on the combination of both of your decisions¡/b¿. I will first explain the activity to you. Then, we will go through a couple of practice rounds to make sure everything is clear. *[Give the earnings table to the participant]*

In this activity each person can choose between two actions: to Cooperate or ¡font color="deeppink"¿Not Cooperate¡/font¿. And each person must choose their action without knowing what the other person chose.

Your Choice	Match's Choice	Your Earnings	Match's Earnings
Cooperate	Cooperate	₦1,000	₦1,000
Cooperate	Not Cooperate	₦500	₦900
Not Cooperate	Cooperate	₦900	₦500
Not Cooperate	Not Cooperate	₦750	₦750

[Enumerators played multiple test games with participants to guarantee comprehension]

You are first going to go through this activity with your match from the Blue Group. We will use the answer the person from the Blue Group gave in this activity when they were matched with another member of the Blue Group.

What do you choose? [Cooperate/Not Cooperate]

Now you are going to go through this activity again, but this time your match will be from the Green Group. We will use the answer the person from the Green Group gave in this activity when they were matched with a person from the Blue Group.

What do you choose? [Cooperate/Not Cooperate]

Now you are going to go through the same activity, but this time the possible earnings for you and your match are going to be a little different. *[Give participants the new table of possible earnings]*

Your Choice	Match's Choice	Your Earnings	Match's Earnings
Cooperate	Cooperate	₦1,000	₦1,000
Cooperate	Not Cooperate	₦500	₦700
Not Cooperate	Cooperate	₦700	₦500
Not Cooperate	Not Cooperate	₦600	₦600

You are first going to go through this new version of the activity with your match from the Blue Group. We will use the answer the person from the Blue Group gave in this activity when they were matched with another member of the Blue Group.

What do you choose? [Cooperate/Not Cooperate]

Now you are going to go through this activity again, but this time your match will be from the Green Group. We will use the answer the person from the Green Group gave in this activity when they were matched with a person from the Blue Group.

What do you choose? [Cooperate/Not Cooperate]

Module 7: Attitudes on Policy

The following questions are about some policy proposals for the city of Jos. Our objective is just to get a sense of the possible support and downsides these policies may have. We don't necessarily think they are good or bad, we just want to get an assessment, so please feel free to answer whatever you feel. Remember, your answers are completely anonymous and will never be analyzed individually.

• Policy 1: New settlements in Jos should mix Christians and Muslims

This policy may have some possible downsides. Tell me if you agree or disagree that the following is a downside of this policy:

Christians and Muslims have different ways of living that simply cannot coexist together
.[Completely Disagree / Somewhat Disagree / Somewhat Agree / Completely Agree]

Some families would not be able to trust their neighbors in these mixed settlements
.[Completely Disagree / Somewhat Disagree / Somewhat Agree / Completely Agree]

• Policy 2: Schools in Jos should have a mix of Christian and Muslim children and teachers

This policy may have some possible downsides. Tell me if you agree or disagree that the following is a downside of this policy.

Christians and Muslims have different ways of educating their children that simply cannot be integrated
.[Completely Disagree / Somewhat Disagree / Somewhat Agree / Completely Agree]

The safety of our children would be at risk in these mixed schools
.[Completely Disagree / Somewhat Disagree / Somewhat Agree / Completely Agree]

• Policy 3: All hospitals in Jos should have a mix of Christian and Muslism doctors seeing both Christian and Muslim patients

This policy may have some possible downsides. Tell me if you agree or disagree that the following is a downside of this policy.

Doctors from our religion should focus only on helping people from our religion, and they would not be able to do this in these mixed hospitals
.[Completely Disagree / Somewhat Disagree / Somewhat Agree / Completely Agree]

People from our religion could not trust all doctors in these mixed hospitals
.[Completely Disagree / Somewhat Disagree / Somewhat Agree / Completely Agree]

• Policy 4: All marketplaces in Jos should have Muslim and Christian sellers and buyers

This policy may have some possible downsides. Tell me if you agree or disagree that the

following is a downside of this policy.

We would be supporting businesses of people that do not belong to our religion, and we should not do this
.[Completely Disagree / Somewhat Disagree / Somewhat Agree / Completely Agree]

People from my religion will not be able to trust all the sellers and products in these mixed markets
.[Completely Disagree / Somewhat Disagree / Somewhat Agree / Completely Agree]

Module 8: Social Desirability Bias

In the following questions of the survey you will be presented situations that reflect possible perceptions about yourself. In each situation tell us if you agree or disagree with the statement.

- I sometimes feel resentful when I don't get my way [Agree/Disagree]

- On a few occasions, I have given up doing something because I thought too little of my ability [Agree/Disagree]

- I am always courteous, even to disagreeable people [Agree/Disagree]

- There have been times when I felt like rebelling against people in authority even though I knew they were right [Agree/Disagree]

- There have been occasions when I took advantage of someone [Agree/Disagree]

- I sometimes try to get even rather than forgive and forget [Agree/Disagree]

- I have never been irked when people expressed ideas very different from my own [Agree/Disagree]

- There have been times when I have been quite jealous of the good fortune of others [Agree/Disagree]

- I am sometimes irritated by people who ask favors of me [Agree/Disagree]

- I have deliberately said something that hurt someone's feelings [Agree/Disagree]

D3. Money Allocation Decisions Algorithm

To recover the social preferences I use a series of money allocation decisions. This ultimately identifies the willingness to pay a person has to either help or hurt their match in a fixed amount. To identify the parameter of interest in the least amount of questions possible, I use an algorithm that picks the next money allocation decision a participant will face depending on the previous decision the participants has made. The algorithm works as follows.

There are 20 money allocation decisions. In 10 of them, picking Option 2 implies increasing the match's payoff in 500 (altruism side). In the other 10, picking Option 2 implies decreasing the match's payoff in 500 (hate side). Within each side, each money allocation decision varies the price for Option 2 from 0 to 450, by jumps of 50.

First, all participants face the same four money allocation decisions. These are the m.a.d. of price 0 and price 50 on the altruism side and the hate side. Using these four answers I used the following rule to set people into one of the two sides for what was left (altruism or hate). Within each side, the next m.a.d. participant got was that of price 250. From this point, depending on the decision, the next m.a.d. presented followed the decision tree showcased below.

Rule to assign people to hate or altruism side:

Rational altruistic
A0: 2 / A50: 1 / H0: 1 / H50: 1
A0: 2 / A50: 2 / H0: 1 / H50: 1

Leans altruistic
A0: 2 / A50: 2 / H0: 1 / H50: 2
A0: 2 / A50: 2 / H0: 2 / H50: 1
A0: 1 / A50: 2 / H0: 1 / H50: 1
A0: 1 / A50: 2 / H0: 2 / H50: 1

Rational hateful
A0: 1 / A50: 1 / H0: 2 / H50: 1
A0: 1 / A50: 1 / H0: 2 / H50: 2

Leans hateful
A0: 2 / A50: 1 / H0: 2 / H50: 2
A0: 1 / A50: 2 / H0: 2 / H50: 2
A0: 1 / A50: 1 / H0: 1 / H50: 2

A0: 2 / A50: 1 / H0: 1 / H50: 2

Rational neutral (assign randomly)
A0: 1 / A50: 1 / H0: 1 / H50: 1

(Assign randomly)
A0: 2 / A50: 2 / H0: 2 / H50: 2
A0: 2 / A50: 1 / H0: 2 / H50: 1
A0: 1 / A50: 2 / H0: 1 / H50: 2

Prices in the graph below:
price=100 → p=2
price=150 → p=3
price=200 → p=4
price=250 → p=5
price=300 → p=6
price=350 → p=7
price=400 → p=8
price=450 → p=9

Price Algorithm in Money Allocation Decisions

p=5

not pay   pay

p=2   p=8

not pay   pay   not pay   pay

p=1   p=4   p=6   p=9

not pay   pay   not pay   pay   not pay   pay   not pay   pay

○   ○   p=3   ○   ○   p=7   ○   ○

not pay   pay   not pay   pay

○   ○   ○   ○