ANTI-SOCIAL NORMS

Leopoldo Fergusson
José-Alberto Guerra
James A. Robinson

Anti-social norms
Leopoldo Fergusson, José-Alberto Guerra, and James A. Robinson
NBER Working Paper No. 32717
July 2024
JEL No. C91,D74,D91,N46

ABSTRACT

Since formal rules can only partially reduce opportunistic behavior, third-party sanctioning to promote fairness is critical to achieving desirable social outcomes. Social norms may underpin such behavior, but they can also undermine it. We study one such norm the "don't be a toad" norm, as it is referred to in Colombia that tells people to mind their own business and not snitch on others. In a set of fairness games where a third party can punish unfair behavior, but players can invoke the "don't be a toad" norm, we find that the mere possibility of invoking this norm completely reverses the benefits of third-party sanctioning to achieve fair social outcomes. We establish this is an anti-social norm in a well-defined sense: most players consider it inappropriate, yet they expect the majority will invoke it. To understand this phenomenon we develop an evolutionary model of endogenous social norm transmission and demonstrate that a payoff advantage from adherence to the norm in social dilemmas, combined with sufficient heterogeneity in the disutility of those who view the norm as inappropriate, can generate the apparent paradox of an anti-social norm in the steady-state equilibrium. We provide further evidence that historical exposure to political violence, which increased the ostracization of snitches, raised sensitivity to this norm.

Leopoldo Fergusson
Universidad de los Andes
Department of Economics
Cra 1 No 18A - 12
Bogotá, Colombia
lfergusson@uniandes.edu.co

José-Alberto Guerra
Universidad de los Andes
Department of Economics
Cra 1 No 18A - 12
Bogotá, Colombia
ja.guerra@uniandes.edu.co

James A. Robinson
University of Chicago
Harris School of Public Policy and
Department of Political Science
1307 East 60th Street
Chicago, Illinois 60637
and NBER
jamesrobinson@uchicago.edu

A online appendix is available at http://www.nber.org/data-appendix/w32717

I Introduction

Recent research on comparative economic development has emphasized its institutional sources (Acemoglu, Johnson & Robinson 2001, Acemoglu & Robinson 2012, LaPorta, Lopez-de Silanes, Shleifer & Vishny 1998, Nunn 2008). While the focus has often been on specific "extractive institutions", such as various forms of labor coercion and slavery, there is also a recognition that institutional differences between poor and rich countries are both formal and informal and include not just written rules, constitutions and laws, but also a complex gamut of social norms and practices. Following North (1990, 3), "institutions are the rules of the game in a society ... the humanly devised constraints that shape human interaction", shaping incentives and thus determining whether societies achieve desirable outcomes such as interpersonal cooperation, political stability, widespread investment, peace, and accountable governments, to name a few. North (1990, 36) went on to note, "a moment's reflection should suggest to us the pervasiveness of informal constraints ... codes of conduct, norms of behavior, and conventions." While an extensive historical and social science literature has conjectured that the nature of economic development is impacted by norms and informal practices (for example Harrison & Huntington (2000)), there has been little empirical research into these norms, how they may influence social equilibria, or where they may originate.

To the extent that norms or informal institutions are part of the explanation for underdevelopment, they must, in some sense be "anti-social", by perhaps enforcing inefficient practices or trapping society into socially undesirable equilibria. We have little evidence about the nature of such equilibria if they exist.

In this paper, we examine the impact of a particular social norm prevalent in Colombia which has the potential to be an anti-social norm; "*no sea sapo*" (literally "don't be a toad").[1] In practice, this norm means "mind your own business" and is used to respond to people attempting to correct socially inappropriate behavior. As we document in section II, versions of this norm are present in many Latin American countries, and it is frequently invoked in Colombia.

We study the impact of this norm in a standard experimental environment —the dictator game with third-party punishment. This is precisely the context to study the implications of "don't be a toad". The norm seems almost specifically designed to break the enforcement mechanisms that facilitate more fair collective and other socially desirable outcomes through third-party punishment. "Don't be a toad" conveys that you should mind your own

---

[1]The literature on Latin American development has echoed the broader literature emphasizing both formal and informal institutions (Coatsworth 1978, Dell 2010, Engerman & Sokoloff 1997).

business and not snitch on other people, even if they are acting in an anti-social (unfair) way. We hypothesize that adhering to such a norm undermines the effectiveness of the type of third-party enforcement studied by Fehr & Fischbacher (2004), Henrich, McElreath, Barr, Ensminger, Barrett, Bolyanatz, Cardenas, Gurven, Gwako, Henrich et al. (2006), and Ensminger & Henrich (2014).[2]

To investigate this hypothesis, we designed and implemented a series of laboratory experiments with university students in Colombia. In the experiments, a first student (the sender) decides the split of a given endowment with a second student (the receiver). The chosen division could range from providing nothing to providing half of their endowment. We also allow a third student to punish the behavior of the first student. In one variation, senders may invoke the "don't be a toad" norm to the third party if they face punishment.

We show that the possibility of invoking this norm, in practice simply sending a message without consequences for anyone's payoff, reduces the propensity of third-party punishment. As a result, allocations deviate from fairness to the point that living in a society where "don't be a toad" can be invoked is equivalent to having no third-party punishment: a society-less society.

We then investigate the players' normative attitudes to "don't be a toad." We find that 80% of players believe it is socially appropriate for third parties to punish unfairness. Moreover, the same proportion finds invoking the "don't be a toad" norm socially inappropriate. Nevertheless, despite these beliefs, initially, 55% of the receivers expect senders to appeal to the norm should they be punished, and eventually, close to a third of senders respond to punishment by invoking the norm at least once.

This case is, therefore, distinct from what Katz & Allport (1931, 1) call "pluralistic ignorance" defined as "a situation where a majority of group members privately reject a norm, but assume (incorrectly) that most others accept it".[3] In our study, receivers and third parties *correctly* predict that most people disapprove of the "don't be a toad" norm. Still, most receivers expect senders to appeal to such a norm if punished.

In light of this, we define an "anti-social norm" as one that most people disapprove of normatively yet expect to be adhered to by a majority in equilibrium. Thus "don't be a toad" is an anti-social norm.[4]

---

[2]Such behavior is also well documented in ethnographic work (see Flannery & Marcus 2014, for many examples).

[3]Brennan, Eriksson, Goodin & Southwood (2013, 182) discusses this as a potent source of the persistence of "bad norms" (see Bicchieri (2005, Chapter 5) for examples and Bursztyn, González & Yanagizawa-Drott (2020) for a recent empirical illustration).

[4]In the terminology of Bicchieri (2016, 19) we study a "descriptive norm" but not a "social norm" (p. 35) since the latter requires that "most people in their reference network believe they ought to conform to it". As we show, this is not the case with "don't be a toad". Cialdini & Trost (1998) give a similar meaning to "descriptive norm" while using the terminology "injunctive norm" for one most people consider socially

To clarify when such an equilibrium is possible, we develop an evolutionary model inspired by Bisin & Verdier (2000), Bisin & Verdier (2001) and Tabellini (2008). We consider a norm that helps to support the socially undesirable outcome in a social dilemma game.[5] In the model, parents may be of two types: one who thinks the norm is appropriate and another who does not. Parents attempt to transmit their preferences to their children and experience utility if their child shares their attitudes toward the norm. However, after socialization, children play the social dilemma game where they can gain by "deviating" to an anti-social action. A child who thinks the norm is appropriate will do this, while a child who thinks it is inappropriate may do so but has to incur a cost (for violating their internalized norm). Consequently, in the social dilemma equilibrium, children who perceive the social norm as appropriate experience a payoff advantage.

We characterize an interior steady-state equilibrium where, first, most people think the norm is inappropriate, but second, most people also expect that a majority will act consistently with the norm (deviate in the social dilemma). For the first part to be possible, the utility gained by a parent who thinks that the norm is inappropriate of having a child who shares their normative view must be sufficiently large relative to the comparable utility from a parent who regards the norm as appropriate and to the payoff advantage of having a child believing the norm is appropriate in the social dilemma. For the second part to hold, a sufficiently large set of people must consider the norm inappropriate but be willing to violate this internalized norm. Intuitively, the cost they incur by violating their internalized norm is smaller than the deviation payoff.[6] The model suggests that an anti-social norm should become more prevalent when the payoff to acting according to it in the social dilemma increases.[7]

How did Colombia become trapped in such an equilibrium? The literature we discuss in Section II suggests that the "don't be a toad" norm originated in Spain and is quite ancient. Nevertheless, we provide historical evidence that its use in Colombia intensified during the 1950s in the context of a civil war known as *La Violencia* (The Violence), where controlling and concealing information became very important (see Kalyvas (2006) for a general argument in the civil war context). Thus, we hypothesize that the norm spread to sanction those who might have revealed information to enemy combatants.

---

appropriate to conform to.

[5]In Appendix A.B, we describe the precise sense in which the dictator game with third-party punishment can considered as a social dilemma.

[6]An analogy might be to the equilibria of dictator games. While people typically regard a $50-50$ split as "fair" or normatively appealing, in equilibrium, dictators typically give less than this since there is a trade-off between normative preferences and material payoffs.

[7]Thus, this payoff advantage increases the steady-state proportion of those who believe such an anti-social norm to be socially appropriate.

To test the effect of *La Violencia* on equilibrium behavior in our experiment, we surveyed participants about the extent to which their grandparents or great-grandparents were impacted by or exposed to *La Violencia*.[8] Since exposure to *La Violencia* is not random, we focus on the interaction between such exposure and the experimentally induced random variation in the "don't be a toad" treatment.

Controlling for level differences in behavior emerging from exposure, these interactions confirm that invoking the social norm has stronger effects on fairness for participants with a *La Violencia* heritage. In fact, senders with *La Violencia* heritage drive the deviation from fairness when they can send the "mind your own business" message. Exposure to *La Violencia* also appears to reduce the extent to which third parties facing the "don't be a toad" message punish senders and to increase the probability that a sanctioned sender sends the message, though these results are less precise and robust (especially concerning sending the message, a decision we observe for a small sample of punished senders). A possible interpretation of these findings is that family experience of *La Violencia* makes the norm more salient. In response, people are more sensitive to it and are more likely to behave anti-socially.

Our paper is related to several contributions other than those discussed above. We make our analysis as comparable as possible to that of Fehr & Fischbacher (2004) where the focus is on deviations from what they call the "distribution norm" - a fair $50-50$ split in this game. This game enables us to study the consequences of "don't be a toad" in an environment where punishment is not motivated by a desire to increase surplus in addition to enforcing a norm.[9]

Other research has focused on aspects of culture or norms that lead to bad collective outcomes. A seminal version of this would be the 'Amoral familism' proposed by Banfield (1958), which can be interpreted as a norm that justifies non-cooperation. Other studies, for example Leonardi, Nanetti & Putnam (1994) and Guiso, Sapienza & Zingales (2016), developed this hypothesis in terms of social capital and linked it to the under-development of the south of Italy and Tabellini (2010) examined this more comparatively within Europe. Social capital is measured by survey answers about trust and the values one would like one's children to have (see Algan & Cahuc (2013) on the evidence connecting trust to economic development). None of this work directly measures social norms as we do or can examine the

---

[8]Empirical evidence suggests that it is plausible that the norms which people's grandparents developed in the 1950s could be inter-generationally transmitted (see Dohmen, Falk, Huffman & Sunde (2012), Bauer, Chytilová & Pertold-Gebicka (2014), Chowdhury, Sutter & Zimmermann (2022) and see House, Kanngiesser, Barrett, Yilmaz, Smith, Sebastian-Enesco, Erut & Silk (2020) for evidence on children learning the social appropriateness of third-party punishment).

[9]In this game, deviations from the distribution norm do not cause inefficiency (at least with linear payoff functions). In many other closely related contexts (like the trust game and many types of social dilemmas), divergence from fairness or non-pro-social behavior do reduce efficiency.

impact of people invoking them. Foster (1965, 1967) developed the idea that people may hold a zero-sum model of the social world leading to collectively sub-optimal behavior, and recent work by Carvalho, Bergeron, Henrich, Nunn & Weigel (2023), Chinoy, Nunn, Sequeira & Stantcheva (2023) has investigated this empirically. La Ferrara (2019) studies a very different type of norm - people's aspirations - and their development consequences.[10] Gulesci, Jindani, La Ferrara, Smerdon, Sulaiman & Young (2023) propose a theory where harmful norms, in the sense of leading to privately costly actions, persist given people's desire to conform with socially prevalent practices. Our approach shares the interest in understanding the persistence of socially undesirable norms. Yet, we unpack the process of norm transmission: instead of assuming a reduced-form process whereby individuals wish to conform to popular actions, we explore the strategic incentives leading to the emergence of prevalent anti-social norms and, specifically, of following norms that one considers inappropriate. Therefore, our main contribution to this literature is to study a specific norm, show the sense in which it is anti-social, develop a model of how such a norm could persist in equilibrium, and provide some evidence about how the norm became so salient in Colombia.

There is less systematic empirical work on the origins of social norms (see Gelfand, Nunn & Gavrilets (2024) for a review).[11] Banfield and Putnam's work argued that the origins of low levels of social capital in Italy were due to the historical incidence of feudalism while Guiso et al. (2016), Tabellini (2010) link it to past political institutions. Nunn & Wantchekon (2011) propose that low levels of trust in Africa are the legacy of the slave trade and the insecurity it created. Several types of norms, concerning honor and the use of violence (Nisbett & Cohen (1996) and Cao, Enke, Falk, Giuliano & Nunn (2021)') and restrictions on women's behavior (Becker (2024)) have been tied to pastoralism. Brennan et al. (2013, 178) argue that one source of "bad norms" is that they may benefit some people at the expense of others. This is potentially relevant to the origins of "don't be a toad" in *La Violencia*, where the norm likely benefited some armed groups. Much of the literature, however, has taken a "functionalist" perspective, arguing that norms emerge when they promote the collective welfare or solve some problem for the group involved (for example, Ullmann-Margalit (1977)). Henrich, Boyd, Bowles, Fehr, Camerer, Gintis et al. (2004) presents evidence that groups that need to cooperate more develop greater pro-sociality, as do those who engage more in market activities and thus face increased risk of opportunism. This approach leaves little scope for explaining how anti-social norms originate and persist. Moreover, it has the feature of studying equilibrium behavior in games such as the dictator game without empirically

---

[10]See Edgerton (2010) for a compilation and discussion of many ethnographic examples of collectively undesirable norms.

[11]However, there are important case studies of specific norms (see the papers in Hechter & Opp (2011), the chapters in Brennan et al. (2013), and Mackie (1996)).

measuring what norm this responds to.

Our findings on the legacy of *La Violencia* contrast with other hypotheses about the impact of violence and civil war on social norms. Bauer, Blattman, Chytilová, Henrich, Miguel & Mitts (2016) present evidence that exposure to violence can increase pro-sociality. Gelfand (2018) and Eriksson, Strimling, Gelfand, Wu, Abernathy, Akotia, Aldashev, Andersson, Andrighetto, Anum et al. (2021) divide societies into those with tight or loose norms. Colombia would be classed as a society with generally loose norms where people easily break and violate rules. In their argument, conflict leads a society, by an evolutionary process, to tighten its rules. We find just the opposite in Colombia, a key likely reason being that rather than external threats, its history of conflict is marked by internal divisions (a distinction not explored in Eriksson et al. (2021)).

Regarding the experimental literature, our notion of an anti-social norm is distinct from the research of Herrmann, Thöni & Gächter (2008), which has focused on anti-social punishment. These researchers show that people are willing to punish players they believe are overly pro-social, even if this is costly. This is different from what we find.

Our work is also different from research examining how players' ability to punish other players influences the play of games. For example, Cinyabuguma, Page & Putterman (2006) show that allowing punishers to be punished reduces anti-social punishment, but introduces a different phenomenon of "perverse" punishment, where those who punished both high and low contributions are themselves punished. The net effect of allowing punishment on efficiency becomes insignificant mostly because free riders are less likely to be punished, so such behavior is reinforced. Research following Nikiforakis (2008) and Balafoutas, Grechenig & Nikiforakis (2014) who allowed individuals to punish those who punished them (players do not know this information in Cinyabuguma et al. (2006)) is closer to our work. This "second-order" punishment leads to behavior that they interpret as retaliation or revenge and unravels the efficiency benefits of first-order punishment (and possibly feuds, see Nikiforakis & Engelmann (2011)).[12]

There are several key differences between our experiment and all of this research. Firstly, we study a real-world example of a social norm actually used rather than a phenomenon generated in the laboratory. Secondly, in our setup, invoking the norm has no implications for anyone's payoff. In the public goods environment used in the above papers, free riding reduces everyone's payoff, giving a material incentive to punish in revenge and perhaps counter-punish to lower the payoff of the person who reduced someone else's payoff. In

---

[12]Denant-Boemont & Noussair (2007) combines these two experiments and finds severely negative effects on welfare. However, neither Kamei & Putterman (2015) nor Fu, Ji, Kamei & Putterman (2017) find that retaliation, even if it takes place, significantly reduces the benefits of allowing for punishment.

our game, while third-party punishment reduces the sender's payoff, invoking "don't be a toad" influences nobody's payoff. Thus, rather than as retaliation to punish the person who has punished the sender, we interpret this as the sender drawing attention to the fact that the third party deviated from the norm. Though many of these and other studies use the language of "norm," they rarely measure this and use it to refer loosely to equilibrium play. For example, if someone chooses a 50-50 split in the dictator game, they are said to follow a fairness norm. In the papers above where retaliation occurs, it is unclear what norm is involved (see Muñoz Herrera & Nikiforakis (2020) for some progress on this). In our study, we empirically elicit peoples' *normative* attitudes towards both the fair distribution $50 - 50$ norm and the "don't be a toad" norm and their expectations of its use. This allows us to characterize the precise sense in which this is a norm.[13]

II   The incidence and roots of "don't be a toad"

Punishing third-party enforcers of social norms and fairness is present in many contexts and countries, even if prevalent in varying degrees. Expressions often summarize the dictum, like "snitches get stitches" in English, "*cagueta morre cedo*" (snitches die early) in Portuguese, "*chi fa la spia non è figlio di Maria*" (whoever spies is not Mary's son) in Italian and "*wie praat die gaat*" (who talks goes) in Dutch, to name a few. Similarly, derogatory terms against third-party enforcers abound. Examples include snitch and rat in English, *X9* or *dedo-duro* in Brazil, *acusa-cristos* in Portugal, and finally "*sapo*" (or toad) in some Spanish-speaking countries, like Chile, Peru, and notably Colombia where we ran our experiment.

While these terms often originate and are more prevalent among criminals condemning and threatening whistle-blowers, they also permeate daily use. Everyday usage frequently extends beyond the typical criminal context to condemning anyone who criticizes a wrong-doer. Such is the case with the Colombian expression "don't be a toad". A newspaper article search for *sapo* and its related verb (*sapear*) in Colombia's leading national newspaper, *El Tiempo*, produced over 200 relevant entries about the expression (between 1982 and 2012).

Figure I presents a word cloud of the newspaper search. These stories reflect how widespread the norm is in Colombia. While some refer to criminal and violent contexts,

---

[13]The absence of direct information on peoples' actual norms can lead to potentially misleading interpretations. For example, punishing very pro-social people is described as "anti-social punishment". Several evolutionary theories have been proposed which could account for it, e.g., by Rand & Nowak (2011) and Sylwester, Herrmann & Bryson (2013). Nevertheless, De Herdt & de Sardan eds. (2015) document ethnographically in many African bureaucracies the presence of a norm they summarize as: "you shouldn't overdo it". Could someone "over-contributing" in the dictator game violate a similar norm and, consequently, be punished? If so, is such punishment anti-social? Potentially, yes, but to judge this, it helps to know exactly the type of "normative" information we elicit in our experiment.

many refer to situations among the general population. The stories reveal a general understanding that denouncing others will be met with disapproval and, in certain contexts, even violence. A football player discussing compliance with their team's rules for healthy habits states he would not want to expose others and become a *sapo*. A bullied boy in school says that one should denounce such behavior, "even if other kids call you a *sapo*". Other stories about violence and bullying in schools share similar experiences and reveal how the norm is socialized from an early age, with one student even killing a peer who denounced his substance abuse in school. The mayor of Barranquilla, a city on the north coast, refuses to share information on those responsible for injuring 17 policemen and 30 people during an eviction because "he is not a *sapo*". A person who did not receive an official reward says he regrets being a *sapo*.

There is a general understanding of the perverse social implications of this norm despite its prevalence and persistence, and there have been political attempts to change this norm and its unfortunate social effects.[14] Several op-eds complain about the culture of "don't be a toad" and worry about its implications; others (much fewer) defend its logic. Another story complains about local culture, noticing how demanding other drivers to behave correctly on the road is met with insults and demands not to be a *sapo*.

– Figure I here –

The term "*sapear*" extends beyond Colombia. "*Sapear*" and "*sapo*" appear in a 1910 dictionary of *coa*, Chilean criminal slang (Cifuentes 1910), with the definition "publicly denounce a robbery" (*designar en público un robo*) and "spy" (*espía*), respectively. Colombian references appear later, so one possibility is that Colombians adopted the Chilean expression. Another possibility is that the expressions emerged independently, where the physical features of toads (a big mouth, a long tongue, a loud croak) naturally suggest the analogy. The work of Wagner (1950) notes, in addition to Chile, early appearances of *sapo* in dictionaries of expressions in Brazil (1922 and 1945) and Argentina (1946). "*Sapear*" also appears, though with more emphasis on spying than denouncing, in Peruvian texts (e.g., Neyra (1977)). A dictionary of Americanisms published in 1997 also lists *sapo* in Central America, including Panama, Nicaragua, and Costa Rica (Richard & Caplán 1997). Later dictionaries add other countries like Honduras, El Salvador, Nicaragua, and Ecuador.[15]

---

[14]Notably, Antanas Mockus, a former mayor of Bogotá, made an effort to institute a national day for praising toads in Colombia, the "croak-tivity day", aiming to highlight the benefits of social sanctioning (see https://caracol.com.co/radio/2003/04/02/nacional/1049234400_035760.html, last accessed February 2024).

[15]Asociación de Academias de la Lengua Española, Diccionario de americanismos, 2010, https://www.asale.org/damer, last accessed September 2022.

The prevalence of the "don't be a toad" norm across Latin America opens the question of whether there is a shared Hispanic cultural predisposition towards disliking third-party enforcers, and several authors have proposed this link (García-Villegas 2017, 2018). An idea along these lines appears in *Don Quijote* (Book I, Chapter XXII): "These poor fellows have done nothing to you; let each answer for his own sins yonder...it is not fitting that honest men should be the instruments of punishment to others, they being therein no way concerned."[16] That "each man should answer for his sins" evokes a biblical expression that might have inspired Don Quijote.

Another potential Spanish root of the norm is the concept of what Thompson (1994, 148), referring to a large historical literature calls "*Obedézcase, pero no se cumpla*" ("to be obeyed, but not to put into effect")". This principle has been dated to the fourteenth century (see González Alonso (1980) and Mackay (1999)) and captured the idea that natural law or customs put constraints on the laws that kings could implement, even if they were passed. It maintained "that legality was to be defended even against the intervention of the king himself, and that the king's orders were to be weighed against external standards of justice and public policy" (Thompson 1994, 149). The principle became particularly important in Spain with the decline of representative institutions like the *Cortes*. It was exported to the Latin American colonies (see García Gallo (1951, 1972)) and, in some interpretations, became a general principle justifying disobedience to the law. It is possibly a short step from that to regarding behavior aimed at enforcing the law as inappropriate, at least during the colonial period.[17]

Whether emerging from the shared Spanish cultural heritage or not, there is no doubt that Latin American attitudes towards norms, norm enforcement, and the role of third-party enforcers are relatively permissive. In Eriksson et al. (2021), Latin American countries, particularly Colombia, stand as places with 'looser' norms. Closely related to our focus, Figure II shows the extent to which people agree with the statement "In this country if someone acts inappropriately, others will strongly disagree": Latin America is the most permissive region and Colombia the most permissive country in the region. Spain, the most permissive in Europe, is more similar to Latin America than its regional neighbors.

---

[16]Translation by John Ormsby, The Project Gutenberg, https://www.gutenberg.org/files/996/996-h/996-h.htm, last accessed September 2022.

[17]García-Villegas (2017, 2018) discusses that idea that the rule of law is considered secondary to perceptions of justice, and thus norms can be broken when considered unfair. He also adds other related mechanisms of cultural influence stemming from colonial history, including: a notion that religion and the family stand above universal legal principles; an extreme form of unrestrained, almost rebellious, individualism; and finally, a history of colonial dominance and later establishment of highly hierarchical societies which exacerbated most of these behaviors. The passage from Don Quijote appears in García-Villegas (2017), with writer Jorge Luis Borges using it to describe Argentina.

In short, several countries have expressions capturing the essence of the "don't be a toad" norm. Latin America is prominent for having a relatively negative perception of third parties who punish socially inappropriate behavior. Within the region, Colombia stands out, making it an ideal context to study the implications of enabling people to punish third-party enforcers.

III   Experimental design

We conducted a series of online laboratory experiments using oTree (Chen, Schonger & Wickens 2016), involving 408 students recruited through the Universidad de Los Andes Mobile Experimental Lab (Uniandes ME-Lab ORSEE database, Greiner 2015). Our experiments exploit a fairness game based on the dictator game with third-party punishment (Fehr & Fischbacher 2004). In this setup, the first student (the sender) makes decisions regarding the distribution of an endowment between themselves and a second passive student (the receiver). This transfer could be fair, a 50-50 split, or deviate from the fair outcome by sharing less than half the endowment. A third student (the third party) can then punish the sender's behavior.

Additionally, we allow punished senders to invoke the don't be a toad norm by sending a "mind your own business" message to the third party. We do not use the colloquial "don't be a toad" expression to avoid an experimenter demand effect that might invalidate our results. Since we are interested in measuring the effect of this anti-social norm on behavior and expectations, using the exact wording "don't be a toad" could lead our subjects to anticipate or guess our intention and modify their behavior, potentially invalidating our empirical analysis. Also, this wording allows for more comparable replications in different contexts and cultures.[18]

III.A   Treatment arms

We introduce the following experimental treatments to examine whether the "don't be a toad" norm affects fairness in Colombia and to understand the underlying mechanisms.[19]

---

[18]To confirm that "mind your business" is similar to "don't be a toad", we asked subjects in the end-line questionnaire to compare this message to sending a "toad" emoticon (used to invoke "don't be a toad" in online chat conversations). Only 2.5% of our subjects considered them to be entirely different messages. The rest found them comparable: 42.7% thought they conveyed the same message, and 46.2% considered them similar, with 38.7% perceiving the emoticon version as softer and 7.5% viewing "mind your own business" as the milder option.

[19]The full-length instructions appear in Appendix B.C.

We follow a between-subject design in which subjects were randomly allocated into one of the treatments.

Dictator Game (DG). In this treatment, two players engage in the Dictator Game (Berg, Dickhaut & McCabe 1995) in the presence of a third player. The sender starts with an endowment of 100 Experimental Tokens (ET),[20] and decides how much to transfer to the endowment-less receiver, with possible amounts of 0, 10, 20, 30, 40, or 50 ET. The third player has a 50ET endowment but no action set, making it entirely up to the sender to determine the endowment distribution: fair (sending 50ET to the receiver) or deviating from fairness (sending strictly less than 50ET).

Third-Party Game (TP-DG). In this treatment, senders and receivers face the same setup as in the Dictator Game, but the third player can now sanction the sender after observing the transfer to the receiver. Following Fehr & Fischbacher (2004), each deduction point imposed by the third party on the sender reduces the third party's payoff by 1ET and the sender's payoff by 3ET.

Third-Party Game + *Sapo* Message (TP-DG-S). Building on the Third-Party Game interaction, this treatment allows senders, when sanctioned by the third party, to invoke the don't be a toad norm without impacting payoffs. Senders can send a "mind your own business" message or no message to the third party at no cost.

Since the senders' behavior in the Dictator Game is the reference point for studying changes in the other two treatments, we introduced the passive third player in this baseline scenario. This ensures that changes in senders' behavior in the other treatments do not reflect the potential effects of the mere presence of a third player.

Furthermore, since the "mind your own business" message is costless to the sender and third parties who receive it, we avoid concerns that counter-punishment motivated by pecuniary factors explains the observed behavior.

---

[20]Equivalent to COP 10,000 (approximately USD 3), about three times the minimum hourly wage or approximately 8.5% their average weekly income.

III.B   Additional details

The described interaction repeats over five decision rounds.[21] We base our main findings on first-round behaviors and expectations since they are unaffected by the evolving unfair behavior, the effectiveness of social sanctions, or changes in using the "mind your own business" message during the experiment. This provides a clean and well-identified baseline influenced solely by the random allocation into treatments.[22] Furthermore, in the first round, subjects are unaware that the setup will repeat. This avoids behavioral changes based on expectations of future play and also allows a direct comparison to the one-shot experiment studied by Fehr & Fischbacher (2004).[23]

In each round, before providing any feedback about behavior, we collect receivers' incentive-compatible empirical expectations about other players' behavior, including expected transfers from senders, expected deduction points imposed by third parties at every possible transfer, and expectations regarding whether sanctioned senders would send a "mind your own business" message. We only elicited receivers' beliefs about senders' and third parties' behaviors to avoid subjects distorting their expressed beliefs to justify their actions. Additionally, we use a simple, easy-to-understand method to incentivize truth-telling when eliciting these empirical expectations: receivers earn 3ET each time their expectations match the majority of players' actions. These two properties are desirable features of methods to elicit beliefs, as suggested by Charness, Gneezy & Rasocha (2021).

Additionally, at the end of the session, we gather incentive-compatible normative expectations from receivers and third parties. Using Krupka & Weber (2013)'s coordination game, participants assess the social appropriateness of various behaviors, including whether it is socially appropriate for a sender to transfer less than 50 ET to a receiver, whether it is socially appropriate that a third party deducts points of a sender who shared less than 50 ET, and whether it is socially appropriate that a sender emits the "mind your own business" message after being charged deduction points. Participants respond on a Likert scale ranging from 'Very socially inappropriate' to 'Very socially appropriate' and get paid 1ET if their answers match the one from a randomly chosen participant. This method exploits a coordination game to elicit shared beliefs about the social appropriateness of different actions. Respon-

---

[21]Player roles remain constant across rounds, with anonymous random group re-matching to prevent opponent-dependent strategic behavior. Players receive feedback about actions and payoffs of the sender and the third party. Only one randomly selected round counts toward the final experimental payment.

[22]Furthermore, while we conducted five rounds of the interaction to explore the dynamics of behavior within the laboratory experimental context, analyzing behavior across rounds would compromise causal inference due to the interdependence of subjects' behaviors stemming from their experiences in previous experimental rounds. A more comprehensive analysis of these dynamics would require a structural behavioral model and is reserved for future research.

[23]We employed the exact instructions and incentives as they did for the Third-Party Game.

dents are incentivized to match the modal response of other subjects while using the shared normative beliefs as a focal point.[24] The identity of the players remains entirely anonymous to participants.

Finally, after completing the experimental session, participants answer a non-incentivized questionnaire regarding their socioeconomic characteristics and individual preferences.

Table I provides an overview of the distribution of participants across treatments and roles. On average, individuals earned 6.5 USD[25] for participating in an hour-and-a-half-long experiment.

– Table I here –

IV   Results

IV.A   The effect of "don't be a toad" on fairness and third-party punishment

In this section, we analyze the effect that the possibility of invoking the "don't be a toad" norm has on the behavior of senders, third parties, and receivers' expectations. Our results confirm that allowing senders to use the "mind your own business" message if punished undermines the effectiveness of third parties intervening to sanction behavior that departs from a fair distribution. Specifically, the possibility of invoking the norm reduces the willingness of third parties to punish unfair behavior and the magnitude of their sanctions and increases the deviation from fair transfers. Receivers correctly believe that the mere possibility of invoking this norm will allow senders to capitalize on selfish behavior, reducing the transfers they make. However, they fail to anticipate that third parties will be less willing to sanction unfair behavior with the introduction of the message.

IV.A.1   Fairness

First, we study the baseline level of fair behavior among senders in the Dictator Game. We then compare it to their behavior when a third party can sanction unfair behavior (in the Third-Party Game) and when senders can invoke the "don't be a toad" norm (in the Third-Party + *Sapo* Game). Figure III depicts the average transfer from senders to receivers by treatment in the first round and the instances in which we observe the senders transferring exactly half of their endowment.

---

[24] Aycinena, Bogliacino & Kimbrough (2023) show that this method outperforms others in the literature.

[25] Equivalent to 18.5% of their average weekly income. In Table A.1 of Appendix A, we provide descriptive statistics of our sample.

– Figure III here –

As similarly found by Fehr & Fischbacher (2004), the unconditional average transfer in the Dictator Game is 25ET. It increases to 32.2 when the protocol allows third parties to sanction unfair behavior,[26] representing an almost 30% increase. When the sender can invoke the "don't be a toad" norm, the average amount sent falls to just 26.7ET. Crucially, we cannot reject the null hypothesis that the average transfer is equal to that in the Dictator Game, consistent with the norm entirely undoing the third party's enforcement capacity.

When we analyze the extent of fair behavior, only 16% of senders transfer half of their endowment to the receivers in the Dictator Game. This proportion more than doubles when the third student can punish unfair behavior. However, this positive effect is almost wholly undone, falling to just over 20%, when the players know the sender can use the "mind your own business" message.

Table II presents coefficients and statistical tests from linear regression models. The dependent variable in the first column is the transfer from senders to receivers. In the second column, it is a variable indicating whether the sender transfers precisely half of their endowment. We add several controls, including whether the participant is female, the academic semester, their guess from a Beauty Contest Game question as a proxy for rationality,[27] socioeconomic stratum as a measure (ranging from 1 to 6) of socioeconomic status (SES),[28] and whether the participant's major is Economics or Business. We report bootstrapped robust standard errors.

– Table II here –

The regression coefficients confirm that when senders face the Third-Party Game, they transfer 8.2ET more to receivers than in the Dictator Game (an increase of 30.9%). When senders know they could invoke the "mind your own business" norm if third parties sanction their behavior, the increase is less than 1.8ET (or under 6.6%), and we cannot reject the null hypothesis that this transfer is equal to the transfer in the Dictator Game. The regression

---

[26]Which is close to the average sent amount in the third-party punishment game found by Henrich et al. (2006) across several societies.

[27]Also known as the Guessing Game, in the beauty contest game participants guess a number from 0 to a 100, and the winner is the one who chooses the number closest to 2/3 of the mean of all the guessed numbers. The game has been used in the experimental literature as a proxy for iterative and strategic thinking (Nagel 1995).

[28]Blanco & Guerra (2020) show the relevance of this SES measure, which builds on the Colombian households' classification into socioeconomic strata at the block-level to target taxes and subsidies for public services. Wealthier households (those living in upper strata 4, 5, and 6) cross-subsidize the poorest residents by paying higher utility bills. Sepúlveda Rico, López Camacho & Gallego Acevedo (2014) show evidence that stratum is strongly positively correlated to household income and expenditure.

results confirm that the mere possibility of senders appealing to the "mind your own business" norm counteracts the gains in fair behavior resulting from the inclusion of social sanctioning.

When we investigate the instances in which the senders transfer half their endowment to the receivers, the percentage of senders who share their endowment evenly when a third party could punish unfair behavior is 20 percentage points (pp) larger than when the third party is passive. However, once senders can invoke the "mind your own business" norm, the likelihood of keeping an equal share of the endowment falls and is statistically indistinguishable from senders' behavior in the Dictator Game.

IV.A.2   Third-party punishment

The change we observe in senders' fairness, triggered by the mere possibility of sending the message, may be attributed to the expectation that third parties would be less willing to sanction unfair behavior. Next, we analyze how introducing the option to invoke the "mind your own business" norm affects the willingness to sanction unfair behavior.

Figure IV displays the average points third parties deduct at each transfer senders make in the first round. We use the strategy method to recover the third party's punishment function, a common practice in these games.[29] Notably, the average sanction decreases as transfers approach a fair split. Additionally, the average punishment is consistently higher in the Third Party Game than when the "mind your own business" message is possible.

– Figure IV here –

When senders cannot invoke the "don't be a toad" norm, the average punishment they receive when sharing nothing is approximately 20ET. However, even though it doesn't affect anyone's payoff, the threat of the "mind your own business" message decreases the average sanction for senders who give nothing by approximately 31%, to just 13.3ET. Note also that anti-social punishment, sanctions on fair senders who share the endowment equally with receivers, is minimal in our sample, not exceeding 1.8ET on average.

We estimate a linear model with third parties' punishment levels as the dependent variable in Table III. The baseline category in the model is the Third Party Game. We include indicator variables for each possible sender transfer, a dummy for the Third Party Game + *Sapo* message treatment, and interaction terms between these variables. Additionally, we incorporate all socio-demographic variables as controls in the regression.

---

[29]If this elicitation method rather than the direct-response method introduces a bias, we may underestimate the third party's willingness to punish unfair behavior, as Brandts & Charness (2011) have shown. This is because in the strategy method, subjects state their decisions in advance, under hypothetical scenarios, leading to less emotionally driven responses rather than making decisions in the heat of the moment, as in the direct-response method.

The constant indicates that the conditional average of deduction points charged by third parties in the Third Party Game is 22.6 ET when senders send nothing to the receiver. Consistent with our previous findings, participants in the game with the "mind your own business" message assign 6.4ET fewer deduction points to these completely selfish senders, a reduction of almost 30% in the intensity of social sanctioning.

Additionally, the coefficients for the various transfer levels confirm that third parties sanction more heavily substantial deviations from a fair distribution than distributions closer to the 50-50 split. For instance, third parties facing the Third Party Game treatment are, on average, willing to sanction senders who transferred 40ET with 7.1 deduction points (computed as the constant plus the coefficient associated with Sent = 40). Instead, they sanction a transferred amount of only 10ET with twice that punishment level (approximately 14ET, the constant plus the dummy associated with Sent = 10).

Finally, deduction points react less to changes in the sent amount when third parties can be called out as toads than when they cannot, as indicated by the positive interactions (significant at the 10% level) between the dummy for the game with "mind your own business" message and the transfer levels.

In short, the comparison between the two "Third Party Game" treatments confirms the relevance of the sender's ability to invoke the "don't be a toad" norm by sending the "mind your own business" payoff-irrelevant message. Third parties impose lower sanctions on senders who deviate from the $50 - 50$ division when they can receive a "mind your own business" message. This, in turn, might explain the result in the previous section, where senders show lower levels of fairness in the Third Party Game + *Sapo* message treatment.

IV.A.3   Empirical expectations

We gathered information on the receivers' expectations about senders' sharing decisions. Additionally, we measured receivers' beliefs regarding third-party punishment. This information is relevant to assess how the possibility of sending a "mind your own business" message influences empirical expectations regarding unfair and sanctioning behavior.

Panel (a) of Figure V presents receivers' expected transfers across the different treatments, which notably deviate from the $50 - 50$ split. We find no significant difference between the receivers' expected amounts in the Third Party Game and those in the Dictator Game. Interestingly, compared to the actual transfers, receivers tend to overestimate the

fair behavior of senders in the Dictator Game. At the same time, their expectations are remarkably accurate in the Third Party Game.

Furthermore, the figure illustrates that receivers anticipate that allowing senders to invoke the don't be a toad norm will enable senders to act more selfishly, resulting in an expected reduction in transfers compared to the Third Party Game. These expectations are close when compared to the actual behavior of the senders in the Third Party with *sapo* message treatment (23ET versus 26.7ET).

On the other hand, in panel (b) of Figure V, we depict the receivers' beliefs in the first round regarding third parties' punishment levels for each transferred amount and treatment. Compared to the actual behavior of third parties, the figure indicates that receivers generally expect larger sanctions in both the simple Third Party Game and with the *sapo* message. However, the receivers fail to anticipate that third parties will be less willing to sanction unfair behavior when sending a "mind your own business" message is possible. Hypothesis tests confirm no significant differences in the expected punishment across both treatments. This discrepancy between expected and actual behavior may stem from receivers' failure to foresee the strategic effect of the "don't be a toad" message on equilibrium play, a result linked to agents' imperfect anticipation of others' strategic adjustments (Dal Bó, Dal Bó & Eyster 2018). This discrepancy may be exacerbated when normative expectations diverge from empirical ones, as discussed in the following section.

IV.B "Don't be a toad" as an anti-social norm

To characterize "don't be a toad" as a social norm, we first establish whether most people believe others would invoke it (related to descriptive norms) and whether most consider it socially appropriate (related to injunctive norms).

In this section, we initially show that a third of the punished senders used the "mind your own business" message at least once. However, the receivers significantly overestimate this prevalence; they believe that more than half of the senders who receive punishment for not evenly splitting the endowment would invoke the "don't be a toad" norm.

Equally important, we show that the majority of senders and receivers believe it is socially appropriate to share the endowment evenly. This result suggests the 50-50 split is an injunctive "distribution norm". They also believe that third parties sanctioning behavior that deviates from this "distribution norm" act in a socially appropriate way. However, we provide evidence that most players believe it is socially inappropriate for senders to invoke the "don't be a toad" norm if sanctioned. This last finding sharply contrasts with the empirical expectations and actual behavior of invoking the "mind your own business" message.

IV.B.1 Empirical expectations and don't be a toad

In Figure VI, we plot the percentage of receivers who believe senders would send the "mind your own business" message should they be sanctioned. The light gray bar depicts this empirical expectation in the first round, which shows approximately 55% of receivers expect most sanctioned senders to use the message. The figure also presents the actual incidence of message usage among sanctioned senders.

– Figure VI here –

Our experimental design employs the traditional direct elicitation method to measure the willingness to use the "mind your own business" message rather than the strategy method. Two reasons motivate this decision. Firstly, as shown by Brandts & Charness (2011), both methods will likely yield significantly different results in complex games such as the one we are studying here, which incorporates third-party punishment and the "don't be a toad" message. Furthermore, Chen & Schonger (2023b) empirically show that strategy method estimates may be biased relative to direct elicitation treatment effects, particularly when there is greater emotional salience in the experiment's actions, which may be relevant for senders sanctioned by third parties. Secondly, Chen & Schonger (2023a) theoretically demonstrate that the strategy and direct methods are not strategically equivalent in games where the utility participants experience does not necessarily equal the received payoffs, as it is likely to occur in our interaction with non-pecuniary costs of following or deviating from social norms. Even though our experimental design decision entails the cost of having fewer observations than if we had used the strategy method, we conclude that the benefits of using the direct method outweigh this cost.

We find that 14% of the senders use the "mind your own business" message in the first round. Eventually, close to 33.3% of senders end up responding to punishment by invoking the norm at least once during the experiment (see Figure A.1 in the Appendix A).[30] Both results suggest that a significant percent of our subjects expect and invoke the "don't be a toad" norm.

IV.B.2 Normative expectations

We now focus on investigating the perceived social appropriateness of participants' behaviors. Figure VII presents the normative expectations of receivers (light gray bars) and third parties (dark gray bars) regarding the studied behaviors.

---

[30]Arguably, as approximately 38.7% of subjects believe that the message "mind your own business" is stronger than sending a toad emoticon conveying "don't be a toad", these percentages are a lower bound estimate of the number of participants willing to invoke the "don't be a toad" norm.

Our findings reveal that more than 60% of both receivers and third parties believe it is socially appropriate for senders to transfer half of their endowment to the receiver. This result suggests that the 50-50 split constitutes an injunctive social distribution norm. Furthermore, over 80% of participants consider it socially appropriate for third parties to sanction behavior that deviates from this "distribution norm". This observation indicates that our subjects also hold injunctive social norms regarding third-party punishment of unfair behavior.

Lastly, the figure provides evidence that more than 80% of receivers and third parties find it socially inappropriate for senders to use the "mind your own business" message when sanctioned. Given that we have established that introducing the "mind your own business" message reduces third-party punishment of unfair behavior and diminishes overall fairness, this result suggests that a majority of our participants perceive themselves to be in an equilibrium they find unappealing. Consequently, this justifies labeling don't be a toad as an anti-social norm.

V Model of intergenerational transmission of anti-social norms

We now theoretically investigate how a society may become trapped in an equilibrium where most people believe and transmit that a norm is socially inappropriate yet anticipate that a majority may invoke it. Specifically, we provide the following definition of an "anti-social norm".

Definition 1. *An "anti-social norm" is a social norm that*

(i) *the majority of people disapprove of normatively, yet*

(ii) *expect the majority to invoke in equilibrium.*

Our theory demonstrates that a payoff advantage resulting from adherence to the norm, combined with sufficient heterogeneity in the disutility experienced by those who view the norm as inappropriate yet invoke it, generates the apparent paradox of an anti-social norm.

We develop an evolutionary model of the endogenous transmission of social norms inspired by Bisin & Verdier (2000, 2001), and Tabellini (2008). Parents exert effort to transmit their normative views to their children (vertical socialization), and children also socialize with their peers (horizontal socialization). A parent whose child has the same normative preferences experiences a "normative payoff". If vertical socialization fails to inherit their trait,

children adopt the normative view from the peer they are matched with during horizontal socialization.

Additionally, during horizontal socialization, children play a social dilemma game. Parents care about the norms their children have and the payoff they experience in this game. We model an anti-social norm as one that encourages socially undesirable behavior in a social dilemma. That is, the norm encourages the Pareto-inefficient action.[31] In the game, children choose an action that can be thought of as "cooperate" or "deviate". Children who regard the norm as socially inappropriate experience disutility if they choose to deviate, while children who believe it is socially appropriate do not.

Our main theoretical result (i.e., Proposition 1) is that there is a culturally heterogeneous steady-state distribution with both normative views co-existing. There are two types of parents: those that deem the norm socially appropriate and those that do not. The critical condition refers to the payoffs of the type who thinks the norm is inappropriate. Intuitively, even though there is a payoff benefit from having a child who believes the norm is appropriate, if parents who regard the norm as inappropriate experience a sufficiently large utility from having a child of the same type, they will exert enough effort to transmit their values with a high probability. This guarantees the existence of an interior steady-state.

Additionally, we establish the conditions under which this social norm becomes an "anti-social" one — where most individuals consider it inappropriate, yet they expect a majority of people to use it in equilibrium. That most individuals consider it inappropriate arises (see Corollary 1) when the normative payoff differential of thinking the social norm is inappropriate (i.e., the utility of a parent who believes the norm is inappropriate from having a child with the same normative preferences minus the comparable utility from a parent who deems the social norm as appropriate) exceeds the payoff advantage of thinking it is an appropriate social norm (i.e., the sum of expected gains in a social dilemma game when matched with others whose normative view is either that the norm is socially inappropriate or appropriate). For most individuals to expect a majority using the norm (see Corollary 2), a sufficiently large set of people must consider the norm inappropriate and be willing to violate this internalized norm because the gain when deviating in the social dilemma exceeds their disutility from doing so.

---

[31]As discussed in the introduction, with linear payoffs, unfair choices are not necessarily inefficient in the dictator game with third-party punishment. This experimental decision was important to isolate the reasons for invoking the norm. Nevertheless, in reality, the use of "don't be a toad" clearly reduces the efficiency of social equilibria, so we believe the social dilemma is appropriate for the theoretical argument.

V.A   Setup

A population of individuals is described by a set of cultural traits regarding the social appropriateness of a norm. A parent with trait $n = a$ thinks that the norm is socially *appropriate*. Instead, a parent with $n = i$ considers it socially *inappropriate*. Society has a unit mass and is characterized by a vector $x = (x_a, x_i)$ with $x_i = 1 - x_a$ where x_a is the proportion who think that the norm is appropriate.

In each period, every parent bears a child. Parents with type n exert effort y_n to increase the probability of successfully transmitting their normative view to their child. If unsuccessful, their child would interact randomly with another member of society and adopt their social norm. Parents with normative view n value the social norm of their child, n', expressively with utility $v(n, n')$. We refer to this payoff as the "normative payoff".

In addition to this standard approach, we introduce a "social dilemma payoff". During horizontal socialization, a child with trait n' is matched with a member of society with trait m. Together, they must decide whether to behave according to the dictates of the norm in a social dilemma and reap individual payoffs from this interaction. The parent, ex-ante, lacks knowledge of the specific match their child will encounter. Consequently, the parent experiences an expected utility equal to $U_S(n' \mid m)$, which depends upon the equilibrium behavior of descendants during the social dilemma. Thus, we assume that the expected social dilemma payoff depends solely on the normative views of the parent's child and their match rather than on the parent's normative view.

V.B   Social dilemma game equilibrium

After two children are matched, they choose whether to behave consistently with the social norm in a social dilemma game.[32] We assume that acting consistently with the social norm creates, in expectation, a larger disutility for those considering it inappropriate (for violating their internalized normative view) than for those who believe it is socially appropriate.

We model this by assuming that children with normative view a never experience a disutility from acting according to the norm ($c_a = 0$), while those with normative view i face a disutility $c_i = c > 0$ with probability p and with complementary probability they also face no disutility. A child of normative view n learns their disutility c_n and that of their match,

---

[32]The interpretation of taking an action which is consistent or not with the norm is game-specific, and our main emphasis is on using the "don't be a toad" message in the third-party game with *sapo* message treatment (see Section A.B). Note, however, that behaving unfairly in the dictator game or not sanctioning unfairness in the third-party game are also potential anti-social norms. We have established experimentally that all of these behaviors are considered socially inappropriate by the majority of subjects, yet a majority in equilibrium expects them. In that sense, they are all examples of invoking an anti-social norm.

c_m, and then decides whether to behave according to the norm (action u, as from to 'use' the norm) or not (d, as from to '*don't use*' the norm).

Therefore, we have the following payoffs for the social dilemma game.

Definition 2. *Payoff structure of the social dilemma game.* *The payoff for a child whose normative view is $n \in \{a, i\}$ and chooses action $z \in \{u, d\}$ while their opponent plays action z' is $\pi_n(z, z')$, given by*

(i) $\pi_n(u, u) = 1 - c_n, \pi_n(d, u) = 0, \pi_n(u, d) = \beta - c_n,$ and $\pi_n(d, d) = \alpha.$

(ii) $c_a = 0$; $c_i = c > 0$ with probability $p \in (0, 1)$, and $c_i = 0$ with probability $1 - p$.

(iii) $2\alpha > \beta > \alpha > c > 1$ and $\alpha > \beta - c$.

These payoffs establish a prisoner's dilemma game-like structure whenever the normative view of both players is a. That is, if players believe the social norm is socially appropriate, acting consistently with the norm is always a profitable deviation, regardless of the opponent's strategy —because in condition *(iii)* we have $\beta > \alpha$. The same happens for players with normative view i who don't experience a disutility from invoking the norm. In contrast, for players who consider the norm socially inappropriate and experience a disutility c, not behaving according to the norm is a strictly dominant strategy —because in condition *(iii)* we have that $c > 1$ and $\alpha > \beta - c$. Also, the cooperative strategy profile where none behave consistently with the norm, (d, d), is Pareto-efficient. It represents the utilitarian social optimum of this game, irrespective of the normative view of the players —because condition *(iii)* states $2\alpha > \beta$.

Denote $z_{nm|c_n, c_m}$ as the strategy for a child with normative view n and disutility c_n whose match has normative view m and disutility c_m. $\pi_n(z_{nm|c_n, c_m}, z_{mn|c_m, c_n})$ is the utility of that child when matched with a member of society with normative view m and disutility c_m under strategy profile $(z_{nm|c_n, c_m}, z_{mn|c_m, c_n})$. Lemma 1 describes the Nash equilibrium in pure strategies profile $(z^*_{nm|c_n, c_m}, z^*_{mn|c_m, c_n})$.

Lemma 1. *The Nash equilibrium of the social dilemma game with payoff structure as in Definition 2 is*

$$z^*_{am|c_a, c_m} = u \text{ for all } m \in \{a, i\}, c_a = 0, c_m \in \{0, c\}, \text{ and}$$

$$z^*_{im|c_i, c_m} = \begin{cases} u & if \quad c_i = 0 \\ d & if \quad c_i = c \end{cases} \text{ for all } m \in \{a, i\}, c_m \in \{0, c\}.$$

Then, the on-the-equilibrium social dilemma payoff that the child with normative view n experiences when matched with a member of society with normative view m from the ex-ante perspective of their parent is

$$U_S(n \mid m) = \sum_{c_n \in \{0,c\}} \sum_{c_m \in \{0,c\}} Prob(c_n) Prob(c_m) \pi_n(z^*_{nm|c_n,c_m}, z^*_{mm|c_m,c_n}).$$

Substituting Nash equilibrium strategies, we get

$$U_S(a \mid a) = 1, \qquad U_S(a \mid i) = 1 + p(\beta - 1),$$
$$U_S(i \mid a) = 1 - p, \quad U_S(i \mid i) = (1 - p)(1 + p(\beta - 1)) + p^2 \alpha.$$

Denote $\Delta U_S(a) \equiv U_S(a \mid a) - U_S(i \mid a)$ as the expected gain in the social dilemma game a child experiences from holding the belief that the norm is socially appropriate, compared to having the opposite normative view, when matched with someone who also deems the norm socially appropriate. Similarly, $\Delta U_S(i) \equiv U_S(a \mid i) - U_S(i \mid i)$ is the expected gain in a social dilemma game when matched instead with someone who considers the norm socially inappropriate. Then, $\Delta U_S(m)$ is the payoff advantage, in the social dilemma game, of having a child aligning with the belief that the norm is socially appropriate when matched with a member who holds normative view $m \in N$.

The social dilemma game from Definition 2 implies that,

$$\Delta U_S(m) > 0 \text{ for any match } m \in N \text{ and } \Delta U_S(i) - \Delta U_S(a) > 0 \text{ as long as } \beta - \alpha > 1.$$

That is, the payoff advantage for someone whose normative view is a is positive for any match as long as the relative gain of acting consistently with the norm when the opponent does not (i.e., $\beta - \alpha$) is greater than the relative gain of acting consistently with the norm when the opponent does it as well (i.e., 1).

Under these conditions, the inequality $\Delta U_S(i) - \Delta U_S(a) > 0$ indicates that the payoff advantage of aligning with the social appropriateness of the norm is greater when the opponent in the social dilemma game considers the norm socially inappropriate rather than when they view it as appropriate.

As we demonstrate below, these two results imply incentives to exert effort in transmitting the appropriateness of the norm within a society where the social dilemma payoffs align accordingly.

V.C   Steady-state of normative views

To understand the evolutionary dynamic, we must first determine parental optimal effort to transmit normative preferences to their children. This effort, denoted y_n, incurs a convex quadratic cost. The parent solves the following decision problem:

$$\max_{y_n} \; y_n \left(v(n,n) + \sum_{m \in N} x_m U_S(n \mid m) \right) + (1 - y_n) \sum_{m \in N} x_m \left(v(n,m) + U_S(m \mid m) \right) - \frac{1}{2} y_n^2. \quad (1)$$

Note that equation (1) implies that a child adopting normative views from a random member of society plays the social dilemma game with that same member of society. To simplify exposition, assume that the parent values when their child shares the same normative view but not otherwise, that is $v(n,n) > 0$ and $v(n,n') = 0$ whenever $n \neq n'$. Then:

Lemma 2. *The optimal interior effort for a parent who considers the social norm socially appropriate is*

$$y_a^* = (1 - x_a) \left[v(a,a) + \Delta U_S(i) \right] \quad (2)$$

while that for parents deeming the social norm socially inappropriate is

$$y_i^* = x_a \left[v(i,i) - \Delta U_S(a) \right]. \quad (3)$$

Proof. The derivative of the objective function in (1) with respect to y_n is strictly positive if and only if $(1 - x_n)(v(n,n) - v(n,n')) + \sum_{m \in N} x_m \left(U_S(n \mid m) - U_S(m \mid m) \right) - y_n > 0$, where $n \neq n'$. As $v(n,n') = 0$, we get the result. $\qquad \square$

Regardless of the parent's normative view, both types invest more effort if the expressive value they place on their children sharing their normative view is higher (i.e., $v(n,n)$). Equation (2) implies that parents who consider the social norm to be appropriate reduce their effort y_a^* to transmit this normative view if a larger share of parents share their same view (i.e., x_a), since horizontal socialization might compensate for the lower effort. This effort instead increases with the advantage in a social dilemma game of their child aligning with the social appropriateness of the norm when matched with a member who holds their parent's opposite normative view (i.e., $\Delta U_S(i)$). The expression and intuition for y_i^* in equation (3) is analogous.

Then,

Proposition 1. *Under a social dilemma such as in Definition 2, every trajectory in the*

interior of the simplex converges to the following state of society:

$$x_a^* = \frac{\Delta U_S(i) + v(a,a)}{(\Delta U_S(i) - \Delta U_S(a)) + (v(i,i) + v(a,a))}$$

if $v(i,i) > \Delta U_S(a)$.

Proof. Following Bisin & Verdier (2001), the evolutionary dynamic of the cultural trait a in society is described by,

$$\dot{x}_a = x_a(1-x_a)(y_a^* - y_i^*) = x_a(1-x_a)(\Delta U_S(i) + (1-x_a)v(a,a) - x_a v(i,i) - x_a(\Delta U_S(i) - \Delta U_S(a))),$$

where the second equality follows from substituting equations (2) and (3). We find the following culturally homogeneous boundary steady states $\{x_a^1 = 1, \ x_a^0 = 0\}$ and a heterogeneous one with $x_a^* = \frac{\Delta U_S(i) + v(a,a)}{(\Delta U_S(i) - \Delta U_S(a)) + (v(i,i) + v(a,a))}$. Given the social dilemma payoff structure, $\Delta U_S(i) + v(a,a) > 0$. Then, for $x_a^* \in (0,1)$ we need $v(i,i) - \Delta U_S(a) > 0$. This condition and the payoff structure of the social dilemma game ensure the stability of the heterogeneous steady state. □

Condition $v(i,i) > \Delta U_S(a)$ says that the normative payoff of a parent who thinks the social norm is inappropriate and has a child with the same normative view is larger than the advantage of having a child who aligns with the opposite normative view in a society of members who think that the social norm is appropriate. This condition ensures that a parent who views the social norm as inappropriate would still exert effort to transmit this normative view to their child (i.e., $y_i^* > 0$ in Lemma 1), even if they live in a society fully populated by members who believe the social norm is appropriate.[33]

V.D   Conditions for an anti-social norm

We can now establish the following result related to the condition *(i)* in our definition of an anti-social norm (Definition 1).

Corollary 1. *In an interior stable steady state x_a^* as the one in Proposition 1*

$$x_i^* > x_a^* \text{ if and only if } v(i,i) - v(a,a) > \Delta U_S(i) + \Delta U_S(a).$$

Corollary 1, which immediately follows from Proposition 1, suggests that whenever the normative payoff differential of thinking the social norm is inappropriate (i.e., $v(i,i) - v(a,a)$)

---

[33]Interestingly, if $\Delta U_S(i) > 0$ the result in Proposition 1 holds even for negative $v(a,a)$. A parent who thinks an anti-social norm is socially appropriate and shares this normative view with their child might experience a negative normative payoff. Despite this, given the social dilemma payoff structure of a society that rewards a descendant more for holding such a normative view, the parent may still be willing to exert effort to instill their child with this anti-social normative view.

is larger than the total social dilemma payoff advantage of thinking it is an appropriate social norm (i.e., $\Delta U_S(i) + \Delta U_S(a)$) we observe a social norm which the majority of people disapprove of normatively (i.e., $x_i^* > x_a^*$).

To establish conditions under which the majority expects the social norm to be invoked in equilibrium (condition *(ii)* in the Definition 1), we need to introduce additional structure to the social dilemma game. Based on Lemma 1, to ensure that the majority expects the norm to be invoked in equilibrium —under rational expectations—, the probability of experiencing a large disutility from invoking the norm for those who deem it inappropriate (i.e., p) must not be too large (see Corollary 2). Intuitively, within those who think the norm is inappropriate, we require a sufficiently large sub-set who nevertheless willingly invokes it because the cost they incur by violating their internalized norm is smaller than the deviation payoff in the social dilemma game.

Corollary 2. *Under the Nash equilibrium in Lemma 1*

$$Prob(z^* = u \mid x_a^*) > \frac{1}{2} \text{ if and only if } p < \frac{1}{2x_i^*}$$

Proof. This occurs because $Prob(u \mid x_a^*) = x_a^* + x_i^*(1 - p)$. □

Thus, under the conditions for Corollaries 1 and 2, an anti-social norm in the sense of Definition 1 emerges.

VI *La Violencia* and don't be a toad

The findings in Section IV prompt us to inquire why the "don't be a toad" norm became so prevalent in Colombia. This section traces the norm's expansion to *La Violencia*, a period of intense bipartisan political violence in the mid-XXth Century following the assassination of Liberal *caudillo* Jorge Eliécer Gaitán in 1948. It also explores whether a family history of exposure to *La Violencia* amplifies the influence of the "don't be a toad" norm within our experimental subjects.

VI.A Historical evidence

The main two political parties, the Liberals and the Conservatives, historically alternated between fighting in elections and wars since their inception in the 1840s and 1850s. *La Violencia* was by far the most intense of the conflicts up to that point, with far more significant loss of life. The period was also a critical turning point for expanding the "don't be a toad" norm in Colombia. Indeed, an influential sociological study of the period Guzmán,

Fals-Borda & Umaña-Luna (1962) even suggests the term "sapo" emerged during *La Violencia*. The study lists the new language "that allowed groups in conflict to understand each other" (p. 216). The entry *Sapear/sapo* appears with the common-day equivalent of "*Delatar/delator*" (to inform/informer).

Figure VIII presents further evidence of the norm's expansion, using the corpus of digitized Spanish material on Google published in or referring to Colombia to graph the relative frequency of the word "*sapear*" (which, unlike the noun "*sapo*" is less likely to appear in other contexts). There is an apparent increase following the 1950s, which coincides with *La Violencia*. Moreover, when examining the appearances before 1950, these are mostly either Optical Character Recognition errors (notably, "*sanear*" or to heal is interpreted as "*sapear*").

– Figure VIII here –

Another indicator of the increased popularity and relevance of the expression is that "*sapear*" and "*sapo*" are recurring entries in dictionaries of Colombian expressions since 1950, but not before.[34] Before *La Violencia*, we have found only one earlier reference to such a term, on a list published in 1938 dedicated to the language spoken by criminals in the Central Penitentiary of Bogotá (Wagner 1950).

In short, Guzmán et al. (1962) are wrong to think of "*sapear/sapo*" as "new language", but correct in identifying its importance during *La Violencia*.[35] Crucially, during *La Violencia*, the "don't be a toad" norm had heightened importance and was violently enforced. Condemning and controlling snitches was vital for the parties in conflict, especially for Liberals opposing the Conservative regime.

The traditional Liberal-Conservative conflict ended after a formal power-sharing deal between the two parties. However, violence and concern about snitches persisted as the conflict gradually morphed into less organized warlordism and disputes between various armed groups. Sánchez & Meertens (1983) describe rewards for snitches set up by the government in 1962, seeking to break the links between peasants and groups of Liberal "*cuadrillas*" of bandits or "*bandoleros*" (Figure IX). Towns protecting these groups were also "punished" by the government, leading in turn to a witch hunt of possible *sapos* by *cuadrillas* and sympathizers. A witness recounts one *cuadrilla* murdering a peasant after accusing him that "If you were a true Liberal, you would not be a *sapo*" (Uribe 1990, pp.164-165).

---

[34]We thank Nancy Rozo from *Instituto Caro y Cuervo* for these revisions of specialized dictionaries, investigating as far back as the late XIXth Century work of Rufino José Cuervo (Cuervo 1907) on language in Bogotá and Spanish America.

[35]It is possible, in principle, that the norm was present and prevalent earlier, perhaps using a different term, but we found no evidence for this in the historical record.

The subsequent persistence of conflict and illegal economies (notably, the drug industry) may have helped to perpetuate the norm in Colombia (see Gambetta (2011) for the utility of such norms in the criminal world). From a historiographical perspective, the term "*sapo*" appears frequently in the examples of Figure I in the context of internal armed conflict and illegality.

– Figure IX here –

By the 1970s, the norm had become entrenched in popular culture and day-to-day interactions. Thus, appeals to the norm appear in literary works and television series about criminality (for example, López López (1980)), but also in the simple interactions between ordinary citizens, as our review of newspaper content of section II revealed.

VI.B   Empirical evidence

If norms are transmitted intergenerationally, descendants of those most affected by *La Violencia* might be more sensitive to the "don't be a toad" condition in our experiment. To investigate this relationship, we administered an additional endline questionnaire to a subsample of our subjects, consisting of 223 students. This questionnaire measured the extent to which participants' grandparents experienced violence during *La Violencia*.[36]

Since our subjects were not alive during *La Violencia*, we first verified their knowledge about the period. Hand-coding their responses to an open-ended question revealed that less than 1.4% of our sample did not know about the period. We then asked several questions about their grandparents' political affiliation (Conservative or Liberal) and exposure to *La Violencia*.

In Table IV, we study how exposure to *La Violencia* affects senders' transfers (columns 1 and 2), social punishment of unfair behavior in the Third Party Game (columns 3 and 4), and "mind your own business" messaging (columns 5 and 6). In the odd-numbered columns, *La Violencia* exposure indicates whether the participant knows that at least one grandparent was a victim of *La Violencia*. In the even-numbered columns, we employ the standardized number of paternal grandparents affiliated with the Liberal Party. The motivation for this second measure is threefold. First, the Conservatives holding national power during *La Violencia* promoted snitching against the opposing Liberals. Second, Liberal guerrillas escaping from the government persisted after *La Violencia* and similarly benefited from controlling *sapos*. Finally, voting rights before and during *La Violencia* were only for males and not

---

[36]See the complete questionnaire in section B.B in the Online Appendix. Descriptive statistics are presented in Tables A.2 in the Appendix and B.4 in the Online Appendix.

for females; arguably, the paternal lineage was the most relevant politically. All regressions include sociodemographic controls, and in columns 3 and 4, we also include fixed effects of all the possible amounts that the sender could transfer to the receiver, and their interactions with treatment indicators.

– Table IV here –

Fairness. Columns 1 and 2 show a positive correlation between *La Violencia* heritage and fairness. This correlation is consistent with the literature documenting positive social capital effects of conflict. But since *La Violencia* exposure is non-random, it could also reflect other differences between families, such as any emerging from places of origin or socioeconomic conditions that influence behavior and expose some families to violence more than others.

Thus, we focus on the interaction between the experimentally induced random variation in the "don't be a toad" treatment and exposure measures. After controlling for level differences emerging from exposure, these interaction terms inform us whether specifically invoking the social norm has stronger effects for participants with a *La Violencia* heritage.[37]

Indeed, the coefficients associated with TP-DG-S×*La Violencia* are large ($-31ET$ and $-15ET$ in columns 1 and 2, respectively) and statistically significant. Columns 1 and 2 indicate that, in the sample with family information on *La Violencia*, sanctioned senders with *La Violencia* heritage drive the transfer decrease when they can send the "mind your own business" message to the third party. In other words, subjects with a family history of exposure to *La Violencia* are more responsive to the "don't be a toad" condition.

Notice finally that the mere introduction of third-party punishment (effectively, introducing a "sapo") appears to reduce transfers for subjects with a heritage of *La Violencia* in column 2, where the large negative coefficient for TP-DG×*La Violencia* is not statistically different to the coefficient for TP-DG-S×*La Violencia*. While our focus is on this latter interaction with the "don't be a toad" treatment, this result might reflect that the norm sufficiently conditions such subjects that resistance towards the third-party punishment dominates its oversight effect.

Third-party punishment. Sending the "mind your own business" message also appears to deter punishment of selfish behavior more for subjects with *La Violencia* heritage. In column 4, where exposure is captured with the standardized number of Liberal grandparents, the TP-DG-S×*La Violencia* coefficient is large and significant, reflecting a decrease of $-4.4ET$

---

[37]When we estimate our basic model without *La Violencia* interactions on this sample, the coefficients are more imprecise given the smaller sample but very similar in magnitude to the baseline. Specifically, the coefficient on Third Party Game (TP-DG) is 6 (compared to 8.2 in the baseline) and the coefficient for Third Party Game + Sapo (TP-DG-S) is 1.4 (1.76 in the baseline).

in punishment. The evidence here is a bit weaker, however, since the interaction is small (-0.15) and not significant when using the indicator for grandparent victimization to measure exposure in column 3. In terms of level effects, Column 3 shows no correlation between *La Violencia* heritage and punishing selfish behavior, while column 4 indicates that third parties with one standard deviation more paternal Liberal grandparents are willing to use 3.5ET more to punish unfairness.

Mind your own business message. We employed a direct method to elicit the senders' willingness to use the message. Thus, we only observe whether punished senders in the first round invoke the norm. Albeit the small sample (only 24 senders), column 5 of Table IV using ancestor victimization to measure *La Violencia* offers evidence that *La Violencia* heritage increases the likelihood of using a message (by 0.35 percentage points, relative to a mean of 16.7%). The correlation when using Liberal grandparents in column 6 is small (coefficient 0.082) and insignificant.

While the results in this section are more speculative and imprecise, they suggest a profound, long-run effect of violent social conflict in sensitizing the population to anti-social norms. Higher family exposure to *La Violencia* makes subjects more responsive to the "don't be a toad" treatment, increasing the likelihood of engaging in anti-social behavior.[38]

Our results add to the recent evidence in economics indicating that collective traumatic events may have long-term effects on behavior, such as investment and entrepreneurial decisions (Ashraf, Bryan, Delfino, Holmes, Iacovone & Pople 2022, Dessi 2008), altruism, trust, and attitudes towards modern medicine (Lowes & Montero 2021, Nunn & Wantchekon 2011, Ramos-Toro 2023). The psychological literature (Alexander et al. 2004) shows that trauma can affect behavior and expectations even if the individual did not experience the traumatic episode themselves but rather if it occurred in a previous generation, as is the case with our sample and *La Violencia* in Colombia. This same literature suggests that the discrepancy between anticipated and actual behavior when cued about the "don't be toad" norm on individuals with a family history of high exposure to *La Violencia* could be explained by limitations in predicting emotional responses (Schachter & Singer 1962), cognitive biases (Tversky & Kahneman 1974), or the context-dependent nature of memory recall (Polyn, Norman & Kahana 2009).

---

[38]We fail to find an effect from *La Violencia* on empirical or normative expectations, as shown in Tables B.1 and B.3 in the Online Appendix.

VII Conclusions

In this paper, we have studied the impact of a real-life social norm, "don't be a toad" or "mind your own business", on the equilibrium play of a dictator game with third-party punishment. Though "don't be a toad" is common in Colombia, close or identical versions are present globally and are particularly salient in various parts of Latin America.

Our key finding is that the mere possibility of invoking the "don't be a toad" norm increases the deviation from fair social outcomes. Subjects are willing to justify selfish behavior by invoking the "don't be a toad" norm. Moreover, players expect others to invoke the norm and anticipate facing fellow players' condemnation for inviting fair behavior.

Despite using the norm, players believe that it is normatively unappealing. In other words, they use it and expect it to be used despite believing it is bad for society. We showed that, theoretically, this counterintuitive combination could arise as an equilibrium of an evolutionary model where, although players find a society that disapproves of "don't be a toad" more desirable, there can be payoff benefits of invoking the norm, particularly when matched to someone who does not approve of it.

"Don't be a toad" fits our theoretical definition of an "anti-social norm": one that most people disapprove of normatively yet expect to be adhered to by a majority in equilibrium. This one is particularly important among the many potential anti-social norms because it is designed to offset the most essential mechanism for social norm enforcement: third-party punishment. In this sense, "don't be a toad" is doubly anti-social. Moreover, the effects we present are quantitatively important, to the point that living in a society where people may invoke the norm is equivalent to living without third-party punishment. In other words, the norm does not just offset but entirely counteracts social norm enforcement through third-party punishment.

We also provided tentative evidence that the roots of the norm's expansion may lie in the 1950s during an intense civil war, *La Violencia*, where insurgent groups and armed forces punished people who betrayed them or potentially gave information or help to opponents. Such a period heightened the payoff advantage of invoking the norm and this effect may have persisted culturally, especially among those most affected by the violence.

Our findings have interesting implications for interventions that counteract undesirable norms. A nascent literature has explored various types of interventions that might be effective at perturbing such norms (see Rhodes, Shulman & McClaran (2020) for an overview). In the case of pluralistic ignorance, this simply involves the transmission of information. In other cases, scholars have experimented with various forms of "contact" and situations where deliberation can take place; see, for example, Lowe (2021) or Webb (2023). In all cases,

the focus is on changing people's normative stances. Our results show that at least in the case we study, people already collectively believe the norm to be undesirable. Therefore, one cannot change the equilibrium by targeting people's normative views. In our model, the critical thing is to change the payoff matrix in the social dilemma.

References

Acemoglu, D., Johnson, S. & Robinson, J. A. (2001), 'The colonial origins of comparative development: An empirical investigation', *American Economic Review* **91**(5), 1369–1401.

Acemoglu, D. & Robinson, J. A. (2012), *Why nations fail: The origins of power, prosperity, and poverty*, Crown Books.

Alexander, J. C. et al. (2004), 'Toward a theory of cultural trauma', *Cultural trauma and collective identity* **76**(4), 620–639.

Algan, Y. & Cahuc, P. (2013), 'Trust and growth', *Annual Review of Economics* **5**, 521–549.

Ashraf, N., Bryan, G., Delfino, A., Holmes, E., Iacovone, L. & Pople, A. (2022), 'Learning to see the world's opportunities: The impact of imagery on entrepreneurial success', *London School of Economics & Political Science, Working Paper* .

Aycinena, D., Bogliacino, F. & Kimbrough, E. O. (2023), 'Measuring norms: A comparison of the predictive and descriptive power of three methods', *Available at SSRN 4663919* .

Balafoutas, L., Grechenig, K. & Nikiforakis, N. (2014), 'The wages of dishonesty: The supply of cheating under high-powered incentives', *Economic Letters* **122**, 308–310.

Banfield, E. (1958), *The moral basis of a backward society*, Free Press paperback, Free Press.

Bauer, M., Blattman, C., Chytilová, J., Henrich, J., Miguel, E. & Mitts, T. (2016), 'Can war foster cooperation?', *Journal of Economic Perspectives* **30**(3), 249–274.

Bauer, M., Chytilová, J. & Pertold-Gebicka, B. (2014), 'Parental background and other-regarding preferences in children', *Experimental Economics* **17**(1), 24–46.

Becker, A. (2024), 'On the origins of restricting women's promiscuity', *Review of Economic Studies* **forthcoming**.

Berg, J., Dickhaut, J. & McCabe, K. (1995), 'Trust, reciprocity, and social history', *Games and Economic Behavior* **10**(1), 122–142.

Bicchieri, C. (2005), *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge University Press, Cambridge University Press.

Bicchieri, C. (2016), *Norms in the Wild: How to diagnose, measure and change social norms*, Oxford University Press, Duke University Press.

Bisin, A. & Verdier, T. (2000), '"Beyond the Melting Pot": Cultural Transmission, Marriage, and the Evolution of Ethnic and Religious Traits', *Quarterly Journal of Economics* **115**(3), 955–988.

Bisin, A. & Verdier, T. (2001), 'The economics of cultural transmission and the dynamics of preferences', *Journal of Economic theory* **97**(2), 298–319.

Blanco, M. & Guerra, J.-A. (2020), 'To segregate, or to discriminate–that is the question: experiment on identity and social preferences', *Documento CEDE* (31).

Brandts, J. & Charness, G. (2011), 'The strategy versus the direct-response method: a first survey of experimental comparisons', *Experimental Economics* **14**, 375–398.

Brennan, G., Eriksson, L., Goodin, R. E. & Southwood, N. (2013), *Explaining Norms*, Oxford University Press.

Bursztyn, L., González, A. & Yanagizawa-Drott, D. (2020), 'Misperceived social norms: Women working outside the home in Saudi Arabia', *American Economic Review* **110**(10), 2997–3029.

Cao, Y., Enke, B., Falk, A., Giuliano, P. & Nunn, N. (2021), 'Herding, warfare, and a culture of honor: Global evidence', *NBER Working Paper* **29250**.

Carvalho, J.-P., Bergeron, A., Henrich, J., Nunn, N. & Weigel, J. L. (2023), Zero-sum thinking, the evolution of effort-suppressing beliefs, and economic development.
URL: *https://www.nber.org/papers/w31663*

Charness, G., Gneezy, U. & Rasocha, V. (2021), 'Experimental methods: Eliciting beliefs', *Journal of Economic Behavior & Organization* **189**, 234–256.

Chen, D. L. & Schonger, M. (2023*a*), 'Invariance of equilibrium to the strategy method i: theory', *Journal of the Economic Science Association* pp. 1–16.

Chen, D. L. & Schonger, M. (2023*b*), 'Invariance of equilibrium to the strategy method ii: experimental evidence', *Journal of the Economic Science Association* pp. 1–13.

Chen, D. L., Schonger, M. & Wickens, C. (2016), 'otree—an open-source platform for laboratory, online, and field experiments', *Journal of Behavioral and Experimental Finance* **9**, 88–97.

Chinoy, S., Nunn, N., Sequeira, S. & Stantcheva, S. (2023), Zero-sum thinking and the roots of U.S. political divides.
URL: *https://www.nber.org/papers/w31668*

Chowdhury, S., Sutter, M. & Zimmermann, K. F. (2022), 'Economic preferences across generations and family clusters: A large-scale experiment in a developing country', *Journal of Political Economy* **130**(9), 000–000.

Cialdini, R. B. & Trost, M. R. (1998), Social influence: Social norms, conformity and compliance, *in* S. T. F. Daniel Gilbert & G. Lindzey, eds, 'The Handbook of Social Psychology, fourth edition', New York: McGraw-Hill, pp. 151–192.

Cifuentes, J. (1910), *Coa jerga de los delincuentes chilenos: estudio y vocabulario*, Imprenta Universitaria.
URL: *https://books.google.es/books?id=4WAwAQAAMAAJ*

Cinyabuguma, M., Page, T. & Putterman, L. (2006), 'Can second-order punishment deter perverse punishment?', *Experimental Economics* **9**, 265–279.

Coatsworth, J. H. (1978), 'Obstacles to economic growth in nineteenth-century Mexico', *The American Historical Review* **83**(1), 80–100.

Cuervo, R. J. (1907), *Apuntaciones críticas sobre el lenguaje bogotano: con frequente referencia al de los países de Hispano-America*, R. Roger y F. Chernoviz.

Dal Bó, E., Dal Bó, P. & Eyster, E. (2018), 'The demand for bad policy when voters underappreciate equilibrium effects', *The Review of Economic Studies* **85**(2), 964–998.

De Herdt, T. & de Sardan eds., J.-P. O. (2015), *Real Governance and Practical Norms in Sub-Saharan Africa: The game of the rules*, Routledge, New York.

Dell, M. (2010), 'The persistent effects of Peru's mining mita', *Econometrica* **78**(6), 1863–1903.

Denant-Boemont, Laurent, D. M. & Noussair, C. N. (2007), 'Punishment, counterpunishment and sanction enforcement in a social dilemma experiment', *Economic Theory* **33**(1), 145–167.

Dessi, R. (2008), 'Collective memory, cultural transmission, and investments', *American Economic Review* **98**(1), 534–560.

Dohmen, T., Falk, A., Huffman, D. & Sunde, U. (2012), 'The intergenerational transmission of risk and trust attitudes', *The Review of Economic Studies* **79**(2), 645–677.

Edgerton, R. B. (2010), *Sick Societies*, Simon and Schuster, New York.

Engerman, S. L. & Sokoloff, K. L. (1997), Factor endowments, institutions and differential paths of growth among the new world economies, *in* S. H. Haber, ed., 'How Latin America Fell Behind', Stanford University Press.

Ensminger, J. & Henrich, J. (2014), *Experimenting with social norms: Fairness and punishment in cross-cultural perspective*, Russell Sage Foundation, New York.

Eriksson, K., Strimling, P., Gelfand, M., Wu, J., Abernathy, J., Akotia, C. S., Aldashev, A., Andersson, P. A., Andrighetto, G., Anum, A. et al. (2021), 'Perceptions of the appropriate response to norm violation in 57 societies', *Nature Communications* **12**(1), 1–11.

Fehr, E. & Fischbacher, U. (2004), 'Third-party punishment and social norms', *Evolution and Human Behavior* **25**(2), 63–87.

Fergusson, L., Ibáñez, A. M. & Riano, J. F. (2020), 'Conflict, educational attainment, and structural transformation: La violencia in Colombia', *Economic Development and Cultural Change* **69**(1), 335–371.

Flannery, K. & Marcus, J. (2014), *The creation of inequality: how our prehistoric ancestors set the stage for monarchy, slavery, and empire*, Harvard University Press, Cambridge.

Foster, G. M. (1965), 'Peasant society and the image of limited good', *American Anthropologist* **67**(2), 293–315.

Foster, G. M. (1967), *Tzintzuntzan: Mexican Peasants in a Changing World*, The Little, Brown Series in Anthropology, Little, Brown.

Fu, T., Ji, Y., Kamei, K. & Putterman, L. (2017), 'Punishment can support cooperation even when punishable', *Economics Letters* **154**, 84–87.

Gambetta, D. (2011), *Codes of the Underworld: How Criminals Communicate*, Princeton University Press.

García Gallo, A. (1951), 'La ley como fuente del derecho en Indias en el siglo xvi', *Annuario de Historia del Derecho Español* **1951**, 607–730.

García Gallo, A. (1972), *Estudios de Historia del Derecho Indiana*, Instituto Nacioal de Estudios Jurídicos.

García-Villegas, M. (2017), *El orden de la libertad*, Fondo de Cultura Económica.

García-Villegas, M. (2018), '¿cómo mejorar a Colombia?', *Bogotá: IEPRI* .

Gelfand, M. (2018), *Rule Makers, Rule Breakers: How Tight and Loose Cultures Wire Our World*, Scribner.

Gelfand, M. J., Nunn, N. & Gavrilets, S. (2024), 'Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change', *Annual Review of Psychology* **18**, 341–378.

González Alonso, B. (1980), 'La fórmula 'obedézcase, pero no se cumpla' en el derecho castellano de la baja edad media', *Annuario de Historia de Derecho* **50**, 469–487.

Greiner, B. (2015), 'Subject pool recruitment procedures: organizing experiments with ORSEE', *Journal of the Economic Science Association* **1**(1), 114–125.

Guiso, L., Sapienza, P. & Zingales, L. (2016), 'Long-term persistence', *Journal of the European Economic Association* **14**(6), 1401–1436.

Gulesci, S., Jindani, S., La Ferrara, E., Smerdon, D., Sulaiman, M. & Young, H. P. (2023), 'A stepping stone approach to norm transitions', *HKS Faculty Research Working Paper Series* **RWP23-013**.

Guzmán, G., Fals-Borda, O. & Umaña-Luna, E. (1962), *La Violencia en Colombia: estudio de un proceso social*, Ediciones Tercer Mundo, Bogotá.

Harrison, L. E. & Huntington, S. P., eds (2000), *Culture Matters: How Values Shape Human Progress*, Basic Books, New York.

Hechter, M. & Opp, K.-D., eds (2011), *Social Norms*, Russell Sage Foundation, New York.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N. et al. (2006), 'Costly punishment across human societies', *Science* **312**(5781), 1767–1770.

Henrich, J. P., Boyd, R., Bowles, S., Fehr, E., Camerer, C., Gintis, H. et al. (2004), *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, Oxford University Press, New York.

Herrmann, B., Thöni, C. & Gächter, S. (2008), 'Antisocial punishment across societies', *Science* **319**(5868), 1362–1367.

House, B. R., Kanngiesser, P., Barrett, H. C., Yilmaz, S., Smith, A. M., Sebastian-Enesco, C., Erut, A. & Silk, J. B. (2020), 'Social norms and cultural diversity in the development of third-party punishment', *Proceedings of the Royal Society B* **287**(1925), 20192794.

Kalyvas, S. (2006), *The Logic of Violence in Civil War*, Cambridge University Press.

Kamei, K. & Putterman, L. (2015), 'In broad daylight: Fuller information and higher-order punishment opportunities can promote cooperation', *Journal of Economic Behavior and Organization* **120**, 145–159.

Katz, D. & Allport, F. H. (1931), *Student Attitudes*, Craftsman.

Krupka, E. L. & Weber, R. A. (2013), 'Identifying social norms using coordination games: Why does dictator game sharing vary?', *Journal of the European Economic Association* **11**(3), 495–524.

La Ferrara, E. (2019), 'Aspirations, social norms, and development', *Journal of the European Economic Association* **17**(6), 1687–1722.

LaPorta, R., Lopez-de Silanes, F., Shleifer, A. & Vishny, R. W. (1998), 'Law and finance', *Journal of Political Economy* **106**(6), 1113–1155.

Leonardi, R., Nanetti, R. Y. & Putnam, R. D. (1994), *Making democracy work: Civic traditions in modern Italy*, Princeton University Press Princeton, NJ, USA.

López López, A. (1980), *El Cartel de los Sapos*, Planeta.

Lowe, M. (2021), 'Types of contact: A field experiment on collaborative and adversarial caste integration', *American Economic Review* **111**(6), 1807–1844.

Lowes, S. & Montero, E. (2021), 'The legacy of colonial medicine in Central Africa', *American Economic Review* **111**(4), 1284–1314.

Mackay, R. (1999), *The Limits of Royal Authority: Resistance and Obedience in Seventeenth-Century Castile*, Cambridge university press.

Mackie, G. (1996), 'Ending footbinding and infibulation: A convention account', *American Sociological Review* **61**(6), 999–1017.

Muñoz Herrera, M. & Nikiforakis, N. (2020), 'Experimental evidence shows that negative motive attribution drives counter-punishment', *Unpublished, NYU Abu Dhabi* .

Nagel, R. (1995), 'Unraveling in guessing games: An experimental study', *The American economic review* **85**(5), 1313–1326.

Neyra, G. (1977), *Argot limeño: o, Jerga criolla del Perú : teoría del argot, argot, jerga y replana, vocabulario y fraseología*, Librería, Importadora, Editora y Distribuidora Lima.
URL: *https://books.google.es/books?id=Lwg8AAAAIAAJ*

Nikiforakis, N. (2008), 'Punishment and counter-punishment in public good games: Can we really govern ourselves?', *Journal of Public Economics* **91**(1-2), 91–112.

Nikiforakis, N. & Engelmann, D. (2011), 'Altruistic punishment and the threat of feuds', *Journal of Economic Behavior and Organization* **78**(1-2), 319–332.

Nisbett, R. E. & Cohen, D. (1996), *Culture Of Honor: The Psychology Of Violence In The South*, Routledge, New York.

North, D. C. (1990), *Institutions, institutional change and economic performance*, Cambridge university press.

Nunn, N. (2008), 'The long-term effects of Africa's slave trades', *Quarterly Journal of Economics* **123**(1), 139–176.

Nunn, N. & Wantchekon, L. (2011), 'The slave trade and the origins of mistrust in Africa', *American Economic Review* **101**(7), 3221–52.

Polyn, S. M., Norman, K. A. & Kahana, M. J. (2009), 'A context maintenance and retrieval model of organizational processes in free recall.', *Psychological review* **116**(1), 129.

Ramos-Toro, D. (2023), 'Social exclusion and social preferences: evidence from Colombia's leper colony', *American Economic Review* **113**(5), 1294–1333.

Rand, D. G. & Nowak, M. A. (2011), 'The evolution of antisocial punishment in optional public goods games', *Nature Communications* **2**, 434.

Rhodes, N., Shulman, H. & McClaran, N. (2020), 'Changing norms: A meta-analytic integration of research on social norms appeals', *Human Communication Research* **46**(2-3), 1–31.

Richard, R. & Caplán, R. (1997), *Diccionario de hispanoamericanismos: no recogidos por la Real Academia:(formas homónimas, polisémicas y otras derivaciones morfosemánticas)*, Cátedra.

Sánchez, G. & Meertens, D. (1983), *Bandoleros, gamonales y campesinos. El caso de la violencia en Colombia*, El Áncora Editores, Bogotá.

Schachter, S. & Singer, J. (1962), 'Cognitive, social, and physiological determinants of emotional state.', *Psychological review* **69**(5), 379.

Sepúlveda Rico, C. E., López Camacho, D. & Gallego Acevedo, J. M. (2014), *Los límites de la estratificación: en busca de alternativas*, Editorial Universidad del Rosario.

Sylwester, K., Herrmann, B. & Bryson, J. J. (2013), 'Homo homini lupus? explaining antisocial punishment.', *Journal of Neuroscience, Psychology, and Economics* **6**(3), 167.

Tabellini, G. (2008), 'The scope of cooperation: Values and incentives', *Quarterly Journal of Economics* **123**(3), 905–950.

Tabellini, G. (2010), 'Culture and institutions: Economic development in the regions of Europe', *Journal of the European Economic Association* **8**(4), 677–716.

Thompson, I. (1994), Castile: Polity, fiscality, and fiscal crisis, *in* P. T. Hoffman & K. Norberg, eds, 'Fiscal Crises, Liberty and Representative Government 1450-1789', Stanford University Press.

Tversky, A. & Kahneman, D. (1974), 'Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty.', *science* **185**(4157), 1124–1131.

Ullmann-Margalit, E. (1977), *The emergence of norms*, Clarendon library of logic and philosophy, Clarendon Press.

Uribe, M. V. (1990), 'Matar, rematar y contramatar', *Revista Controversia* (159-60), 27–203.

Wagner, M. L. (1950), 'Apuntaciones sobre el caló bogotano', *Boletín del Instituto Caro y Cuervo* **1**(2), 181–213.

Webb, D. (2023), 'Silence to solidarity: Using group dynamics to reduce anti-transgender discrimination in india', *https://sites.google.com/view/duncan-webb/home* .

TABLE I. Distribution of participants by treatment and role

Treatment	Senders	Receivers	Third Parties	Total
Dictator Game (DG)	25	26	24	**75**
Third-Party Game (TP-DG)	45	45	44	**134**
Third-Party + *Sapo* Game (TP-DG-S)	66	66	67	**199**
Total	**136**	**137**	**135**	**408**

TABLE II. Linear estimation of the amount sent by the sender and the instances in which senders decide for an equal split to the receiver in the first round, by treatment

Dependent Variable:	Sent Amount	Equality
	(1)	(2)
Third Party Game (TP-DG)	8.230**	0.199*
	(3.383)	(0.104)
Third Party Game + Sapo (TP-DG-S)	1.758	0.061
	(3.312)	(0.090)
Constant	26.598***	0.114
	(7.957)	(0.183)
Controls	✓	✓
Mean Dep. Variable	28.235	0.243
R Squared	0.068	0.056
Participants	136	136
Difference (%): TP-DG vs DG	30.9	174.7
Difference (%): TP-DG-S vs DG	6.6	53.7
p-value H_0: TP-DG = TP-DG-S	0.025	0.107
p-value H_0: TP-DG $\leq$ TP-DG-S	0.014	0.055

Notes: *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Bootstrapped robust standard errors in parentheses. Coefficients come from a linear model where the dependent variable: in column 1, is the amount sent by the sender to the receiver, and in column 2 is a variable indicating when senders decide for an equal split to the receiver. Dictator Game (DG) is the baseline of the analysis. TP-DG is the Third Party Game. TP-DG-S is the TP-DG plus the *Sapo* message "mind your own business". Controls: dummy indicating if female participant, Semester, guess from the Beauty Contest Game Game question as proxy for rationality, Socioeconomic stratum, and a dummy showing if he/she studying economics of business.

TABLE III. Linear estimation of deduction points charged by the third party in the first round

| Dependent Variable:	Deduction Points
	(1)
Sent = 10	-8.591***
	(1.664)
Sent = 20	-11.114***
	(1.879)
Sent = 30	-13.136***
	(2.198)
Sent = 40	-15.523***
	(2.360)
Sent = 50	-17.682***
	(2.552)
Third Party Game + Sapo (TP-DG-S)	-6.402**
	(3.218)
TP-DG-S * Sent = 10	4.576**
	(2.069)
TP-DG-S * Sent = 20	4.606*
	(2.395)
TP-DG-S * Sent = 30	4.733*
	(2.744)
TP-DG-S * Sent = 40	5.269*
	(3.016)
TP-DG-S * Sent = 50	6.159*
	(3.288)
Constant	22.603***
	(3.400)
Controls	✓
Mean Dep. Variable	7.272
R Squared	0.261
Observations	666
Participants	111
p-val H_0: TP-DG-S + TP-DG-S*Sent $\in \{10, \ldots, 50\}$=0	0.488

Notes: *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Bootstrapped robust standard errors clustered by participant in parentheses. Coefficients come from an ordinary least squares regression for the third parties' sample. Dependent variable: is the amount of points a third party would deduct from the sender, given every transfer level. The baseline comparison is Sent=0 and the Third Party Game (TP-DG). TP-DG-S: is the TP-DG plus the *Sapo* message "mind your own business". Controls: dummy indicating if female participant, Semester, guess from a Beauty Contest Game question as proxy for rationality, Socioeconomic stratum, and a dummy showing if he/she studying economics or business. Obs = 111 third parties x 6 decisions = 666.

TABLE IV. Linear estimation of the effect of violence on the amount sent by senders, the deduction points third parties would deduct, and the probability that senders send the "mind your own business" message

|	Sent amount		Deduction Points		Message sent	
	(1)	(2)	(3)	(4)	(5)	(6)
Exposure to *La Violencia*[†]	19.108**	10.297**	0.597	3.472***	0.348*	0.082
	(7.850)	(4.468)	(2.461)	(1.302)	(0.178)	(0.068)
Third Party Game (TP-DG)	10.650	0.054				
	(9.562)	(6.823)				
Third Party Game + Sapo (TP-DG-S)	13.129*	-0.231	-9.137*	-9.746**		
	(7.013)	(5.549)	(4.710)	(4.310)		
TP-DG × *La Violencia*	-15.899	-21.475***				
	(12.986)	(6.781)				
TP-DG-S × *La Violencia*	-30.762***	-15.256***	-0.151	-4.394**		
	(8.042)	(5.296)	(3.368)	(1.808)		
Constant	26.395***	32.739***	20.714***	21.395***	-0.478*	-0.128
	(9.928)	(7.953)	(4.706)	(4.045)	(0.283)	(0.304)
Controls	✓	✓	✓	✓	✓	✓
Mean Dep. Variable	26.232	26.232	7.050	7.050	0.167	0.167
R Squared	0.172	0.207	0.271	0.319	0.400	0.307
Observations	69	69	360	360	24	24
Participants	69	69	60	60	24	24
Treatment samples	(DG,TP-DG,TP-DG-S)		(TP-DG,TP-DG-S)		(TP-DG-S)	

Notes: * p<0.1, ** p<0.05, *** p<0.01. Bootstrapped robust standard errors in parentheses in parentheses. Coefficients come from an ordinary least squares regression. Dependent variable in columns 1 and 2 is the amount sent by the sender to the receiver in the first round, in columns 3 and 4 is the amount of points a third party would deduct, in columns 5 and 6 comes the sanctioned senders' sample in the Third Party plus the *Sapo* message "mind your own business". Sample: Dictator Game (DG) is the baseline of analysis in columns 1 and 2, TP-DG is the Dictator Game with Third Party Punishment and, together with Sent =0, serves as the baseline for the analysis in columns 3 and 4. TP-DG-S is the TP-DG plus the *Sapo* message "mind your own business" is the sample in columns 5 and 6. [†] Each specification uses a different measure of exposure to *La Violencia*: Odd Columns: 1 if the participant knows that at least one grandparent was a victim of *La Violencia*; Even columns: number of paternal liberal grandparents (standardized). Controls: dummy indicating if female participant, Semester, guess from a Beauty Contest Game question as proxy for rationality, Socioeconomic stratum, and a dummy showing if he/she studying economics of business. Columns (3) and (4) additionally control for amount sent dummies and the interaction between these dummies and TP-DG-S.

Figure I. Word cloud map for news stories with the term *Sapo* in Colombia's mass media, 1982-2012
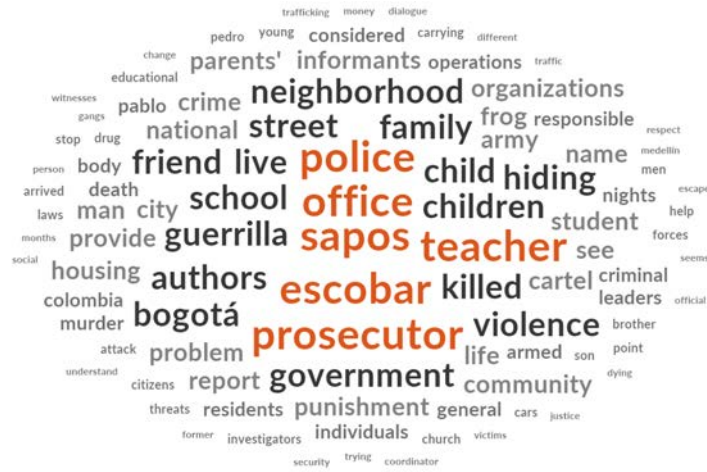

Notes: Selected newspaper entries from *El Tiempo* (1982-2012), Colombia's main national newspaper, including the term "*sapo*". We translated them and introduced them as input files to *NVIVO*, a qualitative data analysis software. We then restricted the analysis to the top 640 most frequently occurring words, with a minimum length of 3 letters to exclude short prepositions, connectors, or other irrelevant words for the analysis (referred to as "stop words" by *NVIVO*) and used the extension grouping method (e.g., treating "lady" or "ladies" as the same) to generate the cloud map.

**Figure II.** Disapproval of inappropriate behavior
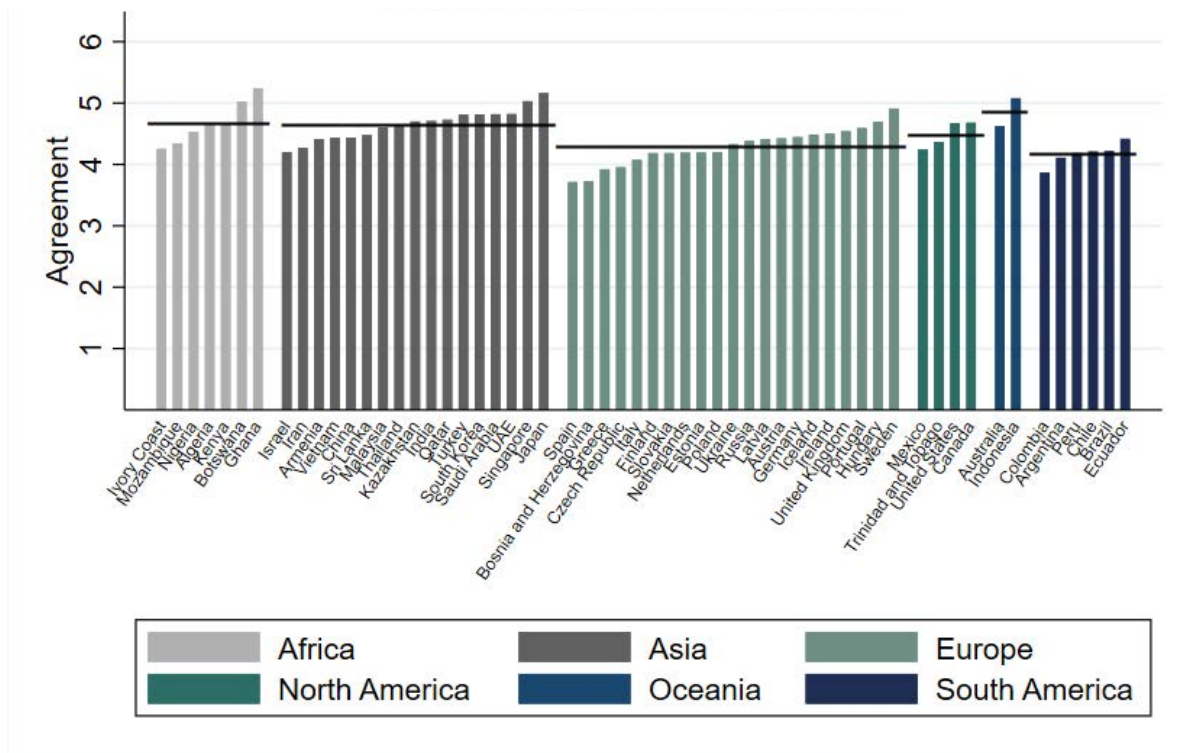
Note: Agreement on a scale of 1 (strongly disagree) to 6 (strongly agree) with the statement "In this country, if someone acts in an inappropriate way, others will strongly disagree". Horizontal bars depict the average agreement across countries from the same region. Source: Eriksson et al. (2021).

Figure III. Average transfer from the senders and instances in which senders decide for an equal split to the receivers, in the first round by treatment

(a) Sent Amount

(b) Equal split
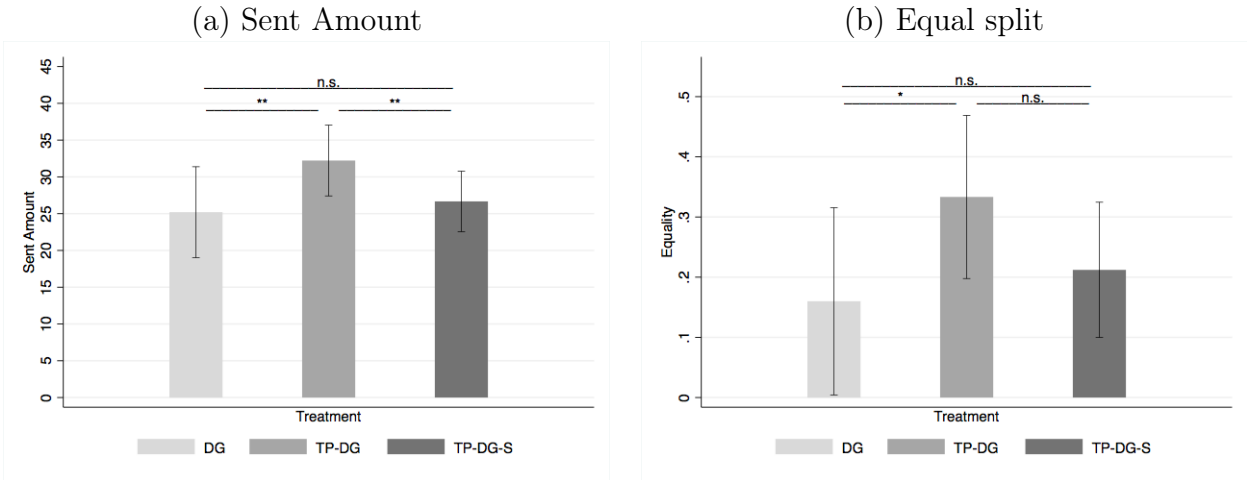
Note: The figure depicts (a) the average amount sent by the sender to the receiver, (b) instances in which senders transfer half their endowment to the receiver depending on treatments in the first round. DG: Dictator Game (light gray) is the baseline of the analysis. TP-DG: is the third-party Game (medium gray). TP-DG-S: is the TP-DG plus the *sapo* message "mind your own business" (dark gray). Dark lines show 95% bootstrapped robust confidence intervals. $***p < 0.01,** p < 0.05,* p < 0.1$, n.s. $p > 0.1$. Reported significance levels come from hypothesis testing from a fully saturated linear model where the dependent variable is the amount sent by the sender to the receiver and regressors are dummy variables associated with treatments.
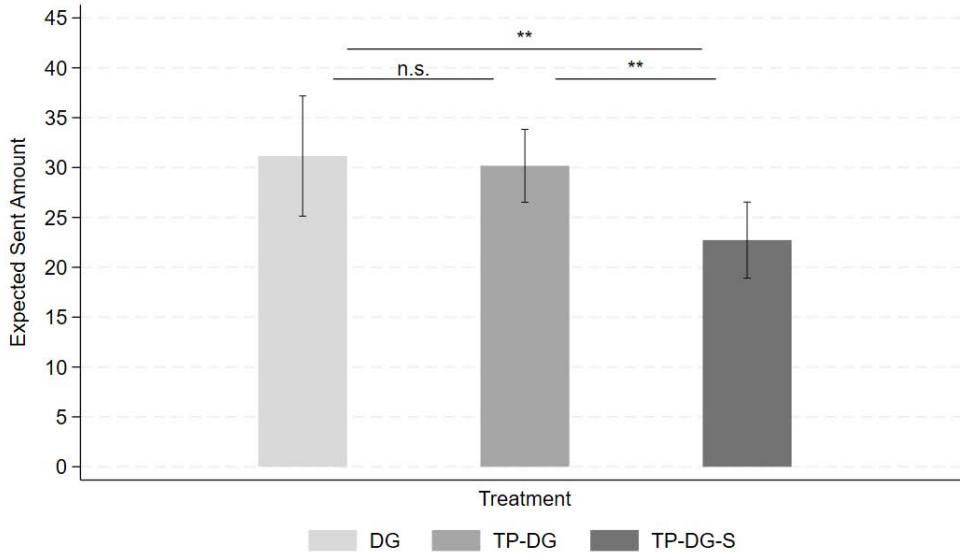
Figure IV. Average deduction points charged by the third parties in the first round, by transfer level and treatment
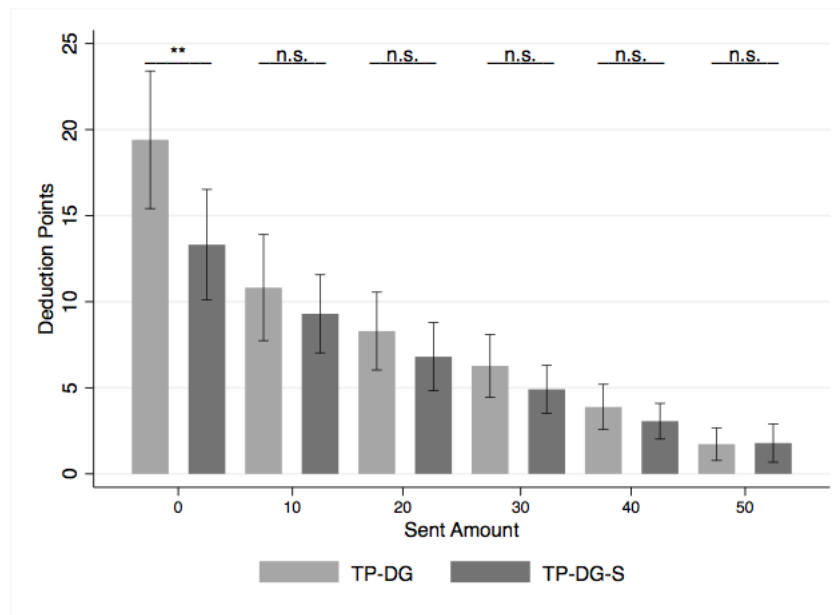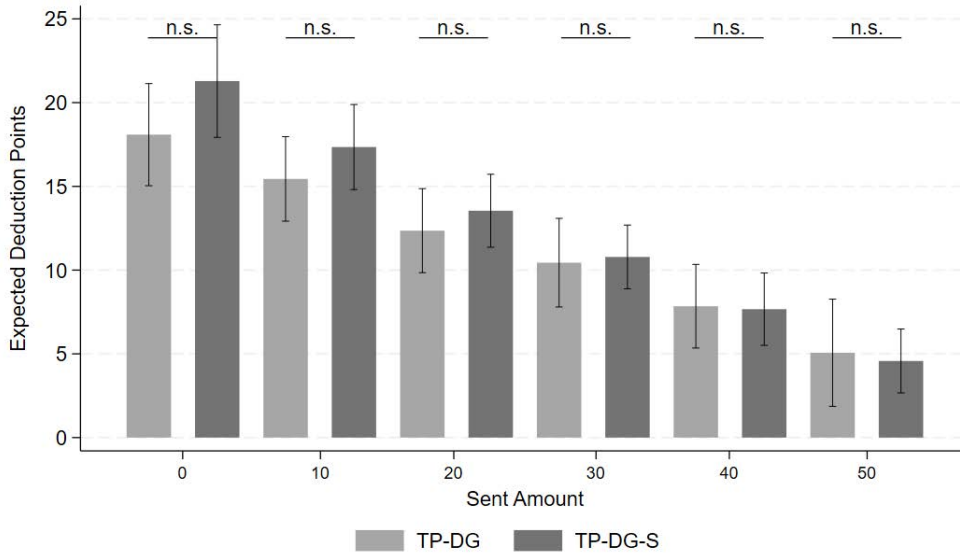

Note: The figure depicts the number of points a third party would deduct from the sender, by each transfer level in the first round. TP-DG: Third Party Game (medium gray). TP-DG-S: Is the TP-DG plus the *Sapo* message "mind your own business" (dark gray). Lines depict 95% bootstrapped robust confidence intervals. $***p < 0.01,** p < 0.05,* p < 0.1$, n.s. $p > 0.1$. Reported significance levels come from hypothesis testing from a fully saturated linear model where the dependent variable is the deduction points charged by the third party and regressors are dummy variables associated with treatments and sent amounts.

45

Figure V. Receivers' expectations about:

(a) Sent amount by senders

(b) Deduction points charged by third parties

Note: The figure depicts the receivers' expectations about: (a) Amount sent by senders, (b) Deduction points by third parties in the first round. TP-DG: Third Party Game. TP-DG-S: Is the TP-DG plus the *Sapo* message "mind your own business". Lines depict 95% bootstrapped robust confidence intervals clustered by participant. ***$p < 0.01$,** $p < 0.05$,* $p < 0.1$, n.s. $p > 0.1$. Reported significance levels come from hypothesis testing from a fully saturated linear model where the dependent variable is (a) the amount sent by the sender to the receiver and regressors are dummy variables associated with treatments, (b) deduction points charged by the third party and regressors are dummy variables associated with treatments and sent amounts

Figure VI. Empirical expectations and actual behavior about invoking don't be a toad norm
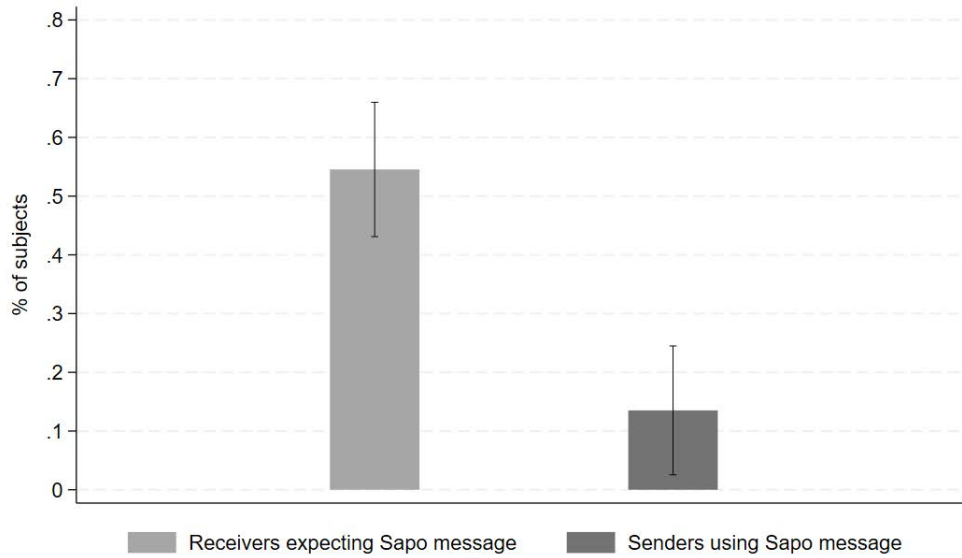
Note: The figure depicts the Percentage of receivers who expect senders to use "Mind your own business" message (light gray) and the Percentage of senders using "Mind your own business" message (dark gray), in the first round. Lines depict 95% bootstrapped robust confidence intervals.

Figure VII. Beliefs about the social appropriateness of being unfair, sanctioning unfairness, and invoking the don't be a toad norm.
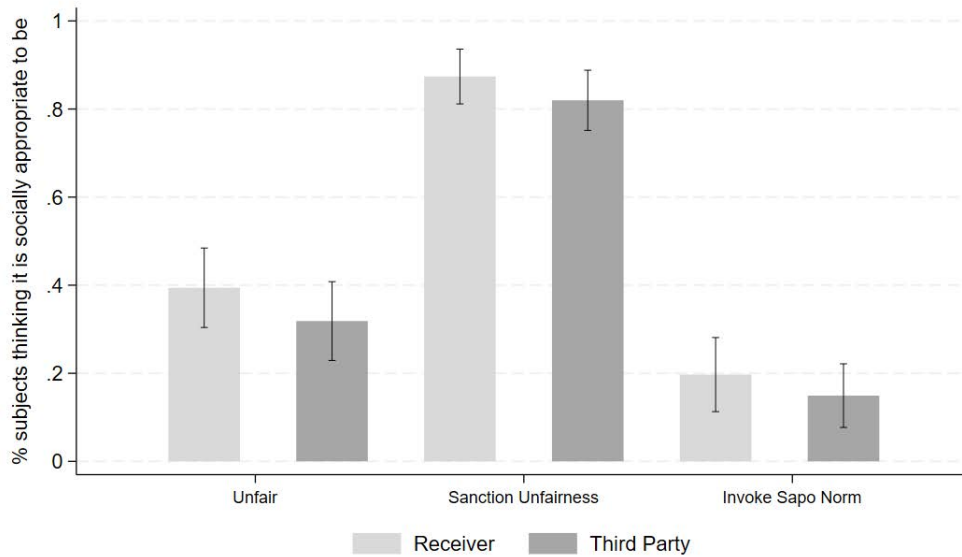
Note: The figure depicts the Percentage of receivers (light gray) and Third parties (dark gray) who answer each of the following questions that they thought the corresponding behavior is "somewhat socially appropriate" or "very socially appropriate". Questions are: (Unfair) Is it socially appropriate if a sender decides to transfer less than 50 points to a receiver?; (Sanction Unfairness) Is it socially appropriate if a third party decides to charge deduction points to the sender if this sender sends less than 50 points to the receiver?; and (Invoke Sapo Norm) Is it socially appropriate if a sender decides to send the message: "Mind your own business" after being charged with deduction points. Lines depict 95% bootstrapped robust confidence intervals.

Figure VIII. Relative frequency of the N-gram "*Sapear*" in Colombian texts, 1900-2005
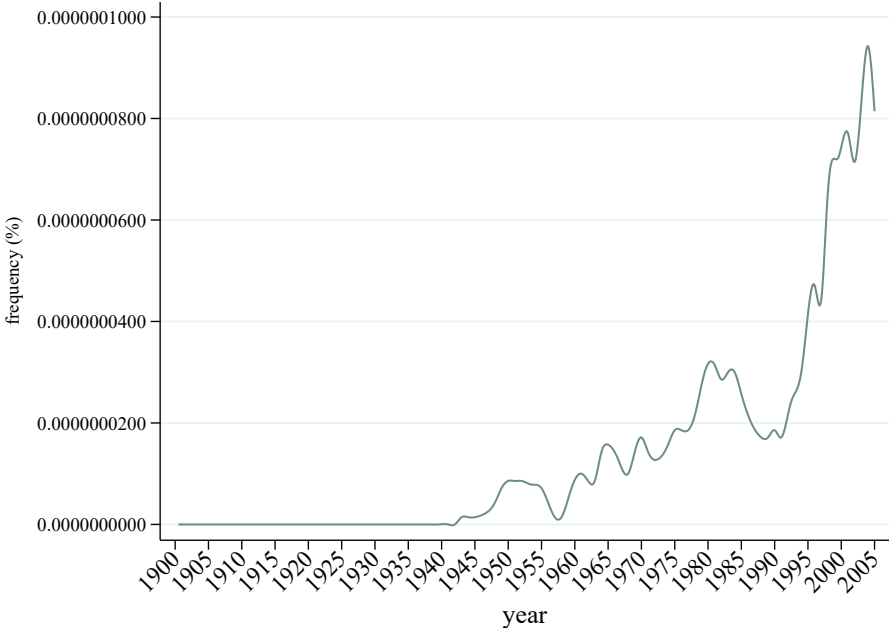
Note: Google's Ngram of "*sapear*" or "*sapiar*" using Google's Spanish corpus of texts, version 2019, published in or referring to Colombia.

Figure IX. Rewards for informants against *bandoleros* during La Violencia

Note: Depictions of posters inviting citizens to denounce leaders of Liberal "*cuadrillas*" in the 1960s in Colombia. These "*sapos*" were in turn sought and punished by *cuadrillas.* Source: Sánchez & Meertens (1983)

A Appendix

A.A Additional Tables and Figures

TABLE A.1. Descriptive statistics and balance tests across treatments: Dictator Game, Third Party Game, Third Party Game + Message

| VARIABLES	Mean **Dictator** Game (75 obs)	Mean **Third Party** Game (134 obs)	Mean **Third Party** Game + **Message** (199 obs)	p-value for H_0 Dictator = TP-DG	Dictator= TP-DG-S	TP-DG= TP-DG-S
Panel A: Participant Variables						
Age	21.31	21.34	21.37	0.93	0.90	0.94
Gender	0.43	0.58	0.55	0.03**	0.074*	0.54
Siblings	1.35	1.51	1.44	0.29	0.48	0.56
Post-graduate	0.13	0.14	0.14	0.86	0.96	0.87
Semester	5.56	5.73	5.21	0.69	0.38	0.11
Experiments before	0.37	0.40	0.49	0.75	0.09*	0.09*
Donations	0.41	0.59	0.55	0.01*	0.05**	0.45
Rationality	36.17	34.89	34.47	0.65	0.54	0.86
Income	106790	121605	154978	0.52	0.18	0.25
Stratum	3.80	3.90	3.87	0.53	0.64	0.87
Panel B: Follow the rules Variables						
Classmates	7.45	7.39	7.34	0.77	0.58	0.78
Politicians	4.23	3.37	3.32	0.01***	0.00***	0.82
Citizens	5.23	5.23	5.11	0.98	0.60	0.51
Family	7.65	7.57	7.65	0.73	0.99	0.65

Notes: In all three treatments there are three player roles: senders, receivers and third parties. Players maintain their role throughout the experiment. No participant participated in more than one treatment. **Experiments Before** is a dummy variable that takes the value of 1 if the participant stated that he/she had participated in an experiment before. **Donations** is a dummy variable that takes the value of 1 if the participant stated that he/she had donated money or had participated as a volunteer with an NGO in the last 12 months. **Rationality** was the answer of the participant to the experiment of *Guess 2/3 of the Average*. Follow the rules panel contains the answer to the following question: "On a scale from 0 (never comply with them) to 10 (always comply with them), how do you think the following groups comply with the laws and social norms in Colombia?". Three participants (one in Dictator Game and two in Third Party Game + Message) reported that they do not know their socioeconomic stratum, and are thus excluded from the descriptive analysis for this variable

Figure A.1. Empirical expectations and actual behavior about use of "mind your own business" across all rounds compared to the first round

(a) Receivers expecting message

(b) Senders using message
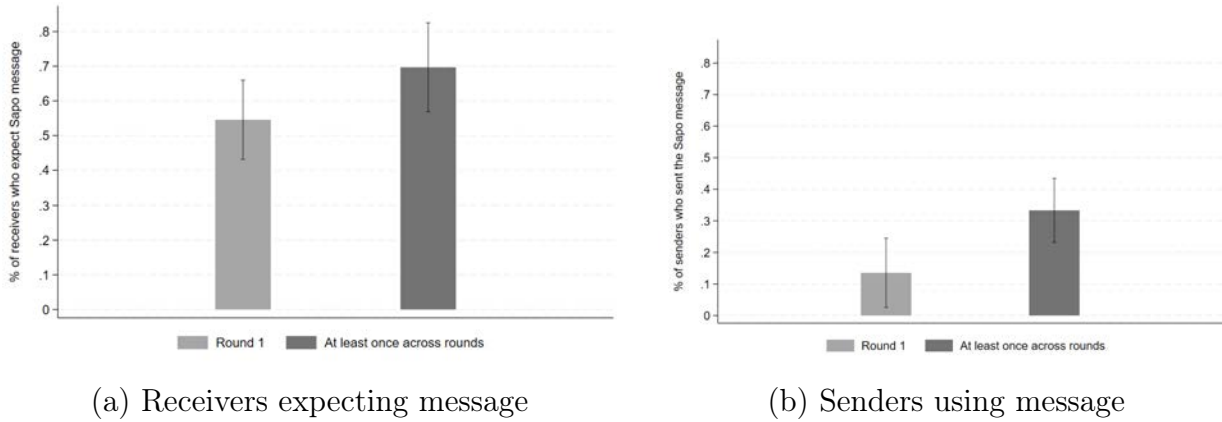
Note: The figure depicts: (a) the percentage of receivers who expect senders to use "mind your own business" message in the first round (light gray) or once across rounds (dark gray); and (b) the Percentage of senders using "mind your own business" message in the first round (light gray) or once across rounds (dark gray). Lines depict 95% bootstrapped robust confidence intervals.

Figure A.2. Receivers' empirical beliefs about invoking the don't be a toad norm by their normative beliefs about social appropriateness of unfairness, sanctioning unfairness, and invoking the don't be a toad norm
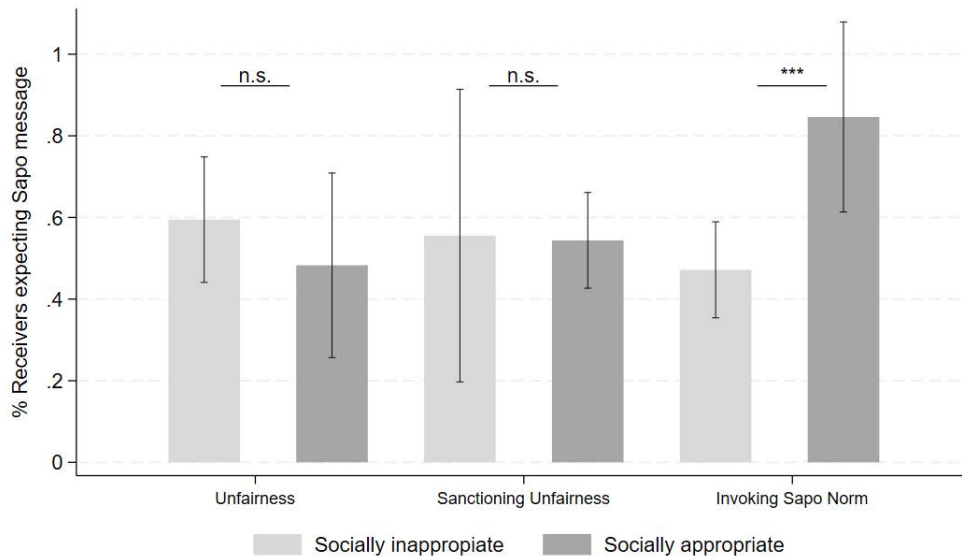
Note: The figure depicts the Percentage of receivers who expect senders to invoke the don't be a toad norm depending on whether they consider that it is socially inappropriate (light gray bars) or socially appropriate (dark gray bars): (Unfairness) If a sender decides to transfer less than 50 points to a receiver?; (Sanctioning Unfairness) If a third party decides to charge deduction points to the sender if this sender sends less than 50 points to the receiver?; and (Invoking Sapo Norm) If a sender decides to send the message: "Mind your own business" after being charged with deduction points. Lines depict 95% bootstrapped robust confidence intervals. $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$, n.s. $p > 0.1$. of the null hypothesis of equality of means.

TABLE A.2. Descriptive statistics of exposure to *La Violencia*

Variables	Observatios	Mean	SD	Min	Median	Max
At least one grandparent affected by *La Violencia*	223	0.52	0.50	0	1	1
Number of liberal paternal great/grandparents	223	0.51	0.74	0	0	2
Relatives of great/grandparents affected by *La Violencia* $(0-10$ scale)	223	5.26	3.05	0	6	10
Relatives of great/grandparents affected by *La Violencia* ($\{0,1\}$ indicator)	223	0.41	0.49	0	0	1
Number of liberal great/grandparents	223	1.16	1.24	0	1	4
Number of liberal maternal great/grandparents	223	0.65	0.77	0	0	2
Family affected by *La Violencia*	223	0.61	0.49	0	1	1
At least one grandparent from a municipality affected by *La Violencia*	223	0.28	0.45	0	0	1

Notes: The previous variables are different measures of exposure to La Violencia: **At least one grandparent affected by La Violencia**: 1 if the participant knows that at least one grandparent was a victim of La Violencia; **Number of liberal paternal great/grandparents**; **Relatives of great/grandparents affected by La Violencia (0-10)**: in a scale from 0 to 10, how were the participant's relatives of grandparents and great grandparents generation affected by La Violencia; **Relatives of great/grandparents affected by La Violencia (0-1)**: 1 if the participant's relatives of grandparents and great grandparents generation were highly affected by La Violencia which is defined as the previous variable being higher than the median; **Number of liberal great/grandparents**; **Number of liberal maternal great/grandparents**; **Family affected by La Violencia**: 1 if their relatives was highly affected by La Violencia according to answers from an open-ended question; **At least one grandparent is from a municipality affected by *La Violencia***, using a measure from Fergusson, Ibáñez & Riano (2020).

A.B Don't be a toad as a social dilemma game

In this section, we translate the situation faced by players in the Third-Party Game + *Sapo* Message (TP-DG-S) treatment into a social dilemma game. Each player's decision-making is shaped by their normative views on whether anti-social norms are inappropriate (i.e., type-i players) or appropriate (i.e., type-a players). Assume the sender (Player $P = 1$) holds normative view n, and the third party (Player $P = 3$) holds normative view m. The sender must decide the amount q_n (where $q_n \in \{0, 0.3\}$) to send from their endowment ($\omega_1 = 1$) to the endow-less receiver (Player $P = 2$). Upon observing this transfer, the third party decides whether to invest t_m units (where $t_m \in \{0, 0.2\}$) out of their endowment ($\omega_3 = 0.5$) to deduct $3t_m$ units from the sender's payoff.[39] After observing t_m, the sender decides whether to invoke the don't be a toad norm by sending the "mind your own business" message ($s_n = 1$) or not ($s_n = 0$). Therefore, the strategy of the sender with normative view n is to choose $z_{1,n} = (q_n, s_n)$ and the third party with normative view m decides over $z_{3,m} = t_m$. The decision tree in Figure A.3 illustrates this interaction. Players' payoffs incorporate material gains, deviations from distribution norms, sanctions for unfair behavior, and the enforcement of social norms. The equilibrium strategies and expected utility differentials reflect how players balance these factors, providing insights into cooperative behavior and norm enforcement.

The intuition of the results, which we develop fully in this section, is the following:

---

[39]The strictly positive values of both q_n and t_m are the observed averages of the sent amount and the deduction points when the sent amount is null in the TP-DG experimental treatment, respectively.

Consider an equilibrium scenario where the sender behaves unfairly ($q_n = 0$), leading the third party to punish them ($t_m = 0.2$), prompting the sender to respond with the *sapo* message ($s_n = 1$). The modeling of this effect in payoffs reveals that sending the message is socially inefficient: both players incur disutility, and the sender primarily sees it as optimal because it mitigates the negative effects of punishment. Now, consider an alternative strategy profile where the sender refrains from sending the *sapo* message ($s_n = 0$). In this scenario, the sender can only achieve a better outcome if the third party reduces their punishment ($t_m = 0$), prompting the sender to initially transfer more ($q_n = 0.3$). Although this profile is not subgame perfect if senders' normative views are $n = a$, it potentially offers Pareto improvement (even benefiting the receiver) if they were to hold normative views $n = i$. However, backward induction indicates that the sender of type $n = a$ finds it optimal to send the *sapo* message, leading to inefficiencies in the more efficient profile that is not a subgame perfect Nash equilibrium. This analysis guides our exploration towards understanding the strategic effects and inefficiencies inherent in norm enforcement and cooperation in the TP-DG-S treatment.

Figure A.3. Decision tree subjects face in the Third-Party Game with *sapo* message
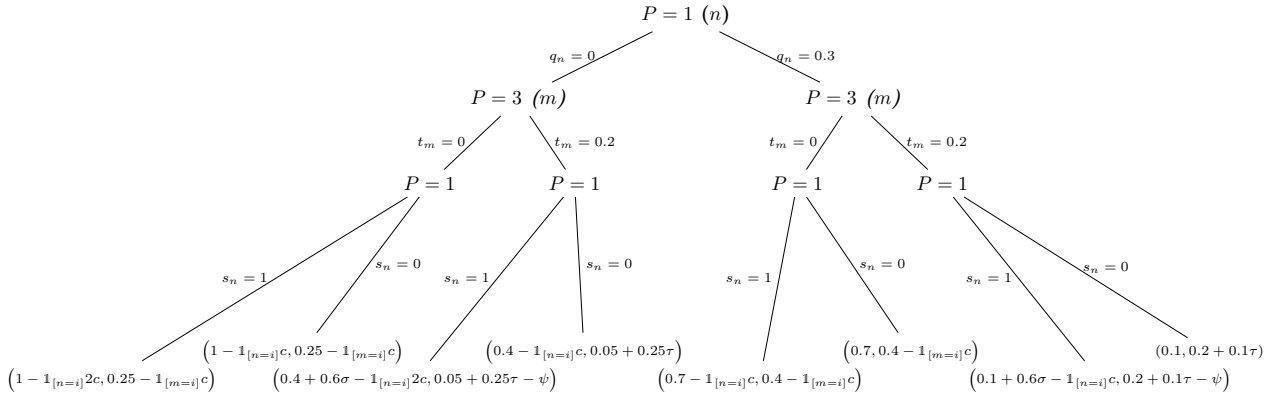
To determine players' payoffs, we assume that subjects care not only about their material payoff but also about deviations from the distribution norm (i.e., the 50-50 split), not sanctioning unfair behavior (i.e., deducting points if the sender behaves selfishly), and when people invoke the don't be a toad norm. In that sense, we define the following preferences

for the senders and third parties in the game,

$$
\pi_{1,n}(z_{1,n}, z_{3,m}) = \overbrace{\omega_1 - q_n - 3t_m(1 - \overbrace{\sigma s_n}^{\text{sapo norm payoff}})}^{\text{material payoff}} - \underbrace{\left(\mathbb{1}_{[q_n=0]} + s_n \mathbb{1}_{[t_m>0]}\right)\mathbb{1}_{[n=i]}c}_{\text{Normative view disutility}}
$$

$$
\pi_{3,m}(z_{1,n}, z_{3,m}) = \overbrace{\omega_3 - t_m}^{\text{material payoff}} - \overbrace{\frac{1}{2}|0.5 - q_n|}^{\text{distribution norm}} \cdot \overbrace{\left(1 - \tau\mathbb{1}_{[t_m>0]}\right)}^{\text{sanction norm payoff}} - \overbrace{\psi s_n \mathbb{1}_{[t_m>0]}}^{\text{sapo norm payoff}} - \underbrace{\mathbb{1}_{[t_m=0]}\mathbb{1}_{[m=i]}c.}_{\text{Normative view disutility}}
$$

Note that $\pi_{1,n}$ implies that senders with normative view n are concerned about their material payoff, which depends on their endowment (ω_1), the amount they decide to transfer to the receiver (q_n), and the points deducted by third parties ($3t_m$). However, senders could invoke the don't be a toad norm ($s_n = 1$) and partially unwind the experienced negative effect from third-party sanctions (by $\sigma \in (0.5, 1]$). Additionally, we assume that senders who believe that anti-social norms are inappropriate (i.e., if $n = i$) would experience disutility c if they transfer nothing to the receiver ($\mathbb{1}_{[q_n=0]}$), or if they invoke the don't be a toad norm when sanctioned ($s_n \mathbb{1}_{[t_m>0]}$).

Now, $\pi_{3,m}$ suggests that third parties care about their payoff after deciding their deduction points ($\omega_3 - t_m$). Additionally, deviations from the 50-50 split reduce their experienced utility by $\frac{1}{2}|0.5 - q_n|$, unless they sanction the sender who deviates from the distribution norm (i.e., $t_m > 0$) which partially alleviates this utility loss (by $\tau \in (0.5, 1]$). They also experience disutility ψ for being called *sapos* after sanctioning selfish behavior. Finally, third parties with normative view $m = i$ incur a disutility of c if they refrain from sanctioning unfair behavior ($\mathbb{1}_{[t_m=0]}$).

For completeness in describing the game, assume the utility of receivers, regardless of their normative views, is defined as $\pi_2(z_{1,n}, z_{3,m}) = q_n - \frac{1}{2}|0.5 - q_n| - \psi s_n + \tau\mathbb{1}_{[t_m>0]}\mathbb{1}_{[q_n=0]}$. The first term is associated with their material payoff, the second with their normative view regarding equality, the third with the don't be a toad norm-associated payoff, and the last term is the payoff they experience from observing that third parties follow the injunctive norm regarding sanctioning unfairness whenever a sender has deviated drastically from the distribution norm.

Let us assume:

$$
0.3 \leq c \leq 0.6\sigma \text{ and } c - (0.2 - 0.1\tau) \leq \psi \leq c - (0.2 - 0.25\tau).
$$

That is, the disutility of invoking a norm for those who deem it socially inappropriate (c) is not too small compared to the material gain of behaving selfishly (i.e, 0.3) nor too high compared to the gain from invoking the don't be a toad norm if sanctioned (i.e., 0.6σ). Additionally, the psychological cost of being called a *sapo* (ψ) is neither too small nor too high compared to the disutility of invoking an anti-social norm (i.e., c) relative to the net gain from sanctioning unfair behavior (which is a function of the deduction points, 0.2, of τ, and the potential losses from experiencing deviations from the 50-50 split – which could be 0.1 if the sender transfers $q_n = 0.3$, or 0.25 if transfers nothing).

Denote $\left(z_{1,nm}^*, z_{3,mn}^*\right) = \left((q_{nm}^*, s_{nm}^*), t_{mn}^*\right)$ as the subgame perfect Nash equilibrium strategies of senders and third parties with normative view n and m, respectively.[40] The following table presents, in the first column, all possible combinations of normative views for the sender (n) and third party (m). The next two columns include the on-the-equilibrium strategies of both senders and third parties. The on-the-equilibrium utilities of senders, receivers, and third parties are depicted in the last three columns of the table (i.e., $U_{S,1}(n \mid m), U_{S,2}(n,m), U_{S,3}(m \mid n)$).

	$z_{nm}^* = \left(z_{1,nm}^*, z_{3,mn}^*\right)$		$U_S(\cdot \mid \cdot)$		
n,m	(q_{nm}^*, s_{nm}^*)	t_{mn}^*	$U_{S,1}(n \mid m)$	$U_{S,2}(n,m)$	$U_{S,3}(m \mid n)$
a,a	$(0,1)$	0	1	$-0.25 - \psi$	0.25
i,a	$(0.3,0)$	0	0.7	0.2	0.4
a,i	$(0,1)$	0.2	$0.4 + 0.6\sigma$	$-0.25 - \psi - \tau$	$0.05 + 0.25\tau - \psi$
i,i	$(0.3,0)$	0	0.7	0.2	$0.4 - c$

Note that only senders who perceive anti-social norms as socially appropriate send a null transfer to the receiver ($q_{am}^* = 0$) and would invoke the "mind your own business" message ($s_{am}^* = 1$), irrespective of the normative view held by the third parties. Conversely, upon observing a null transfer, only third parties who consider completely selfish transfers as inappropriate are willing to charge strictly positive deduction points. In equilibrium, this occurs only when a third party of type $m = i$ is matched with a sender whose type is $n = a$. If both senders and third parties regard anti-social norms as socially inappropriate, or if only the senders think likewise, the senders would transfer 0.3 of their endowment to the receiver, the third party won't deduct points from the sender, and the sender won't invoke the don't be a toad norm.

This structure of payoffs and equilibrium strategies gives us the following expected utility

---

[40]Strictly speaking, by backward induction, t_{mn}^* is a function of s_{nm}^* and q_{nm}^* is a function of t_{mn}^*.

differentials

$$\begin{aligned}
\Delta U_{S,1}(a) &= 1 - 0.7 \geq 0 \\
\Delta U_{S,1}(i) &= 0.6\sigma - 0.3 \geq 0 \\
\Delta U_{S,3}(a) &= 0.2 - 0.25\tau + \psi \geq 0 \\
\Delta U_{S,3}(i) &= c \geq 0.
\end{aligned}$$

Additionally, we have

$$\Delta U_{S,1}(i) - \Delta U_{S,1}(a) \geq 0 \text{ when } \sigma = 1, \text{ and}$$
$$\Delta U_{S,3}(i) - \Delta U_{S,3}(a) \geq 0.$$

Which are consistent with the conditions necessary for results in Proposition 1 to hold.