THE ECONOMIC JOURNAL

The Economic Journal, **133** (*May*), 1626–1640 https://doi.org/10.1093/ej/uead005
Advance Access Publication Date: 18 January 2023

SHIFTING PUNISHMENT ONTO MINORITIES: EXPERIMENTAL EVIDENCE OF SCAPEGOATING*

Michal Bauer, Jana Cahlíková, Julie Chytilová, Gérard Roland and Tomáš Želinský

Do members of a majority group systematically shift punishment onto innocent members of an ethnic minority? We introduce an experimental paradigm, punishing the scapegoat game, to measure how injustice affecting a member of one's own group shapes punishment of an unrelated bystander. When no harm is done, we find no evidence of discrimination against the ethnic minority (Roma people in Slovakia). In contrast, when a member of one's own group is harmed, the punishment 'passed' onto innocent individuals more than doubles when they are from the minority, as compared to when they are from the dominant group.

Scapegoating refers to a social phenomenon where people who feel aggrieved take revenge on another, innocent person. According to social psychology, scapegoating occurs when punishment of the true source of the anger is inhibited and people shift their aggression towards other individuals (see, e.g., the seminal works of Doob *et al.*, 1939 and Allport, 1954). It is often suggested that people are more prone to engage in scapegoating when they can displace aggression on vulnerable and negatively stereotyped minority groups (Bettelheim and Janowitz, 1950; Allport, 1954; Marcus-Newhall *et al.*, 2000). Scapegoating against vulnerable groups is thought to lead to bursts of violence such as lynching, pogroms or even genocide. Scapegoating violates a fundamental fairness principle (Kant, 1965 and classical philosophers) embedded in the legal codes of most modern societies, i.e., that people should be punished only for wrongs they are responsible for and that they intentionally committed. Furthermore, it may also drag minorities into violent conflicts that are completely unrelated to their behaviour and transform individualised tensions into group conflicts.[1] Nevertheless, experimental evidence on how identity of the target shapes people's desire to engage in scapegoating is missing.

* Corresponding author: Michal Bauer, CERGE-EI (a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences), Politických vězňů 7, 111 21 Prague, Czech Republic. Email: bauer@cerge-ei.cz

This paper was received on 28 June 2021 and accepted on 10 January 2023. The Editor was Steffen Huck.

The data and codes for this paper are available on the Journal repository. They were checked for their ability to reproduce the results presented in the paper. The replication package for this paper is available at the following address: https://doi.org/10.5281/zenodo.7404514.

We thank Alexander Cappelen, Ernst Fehr, Michael Kosfeld, Filip Matějka, Matthias Sutter, Bertil Tungodden, seminar participants at the University of Munich, Max Planck Institute in Bonn, UC Berkeley, University of Edinburgh, Goethe University in Frankfurt, Norwegian School of Economics, and participants at the ECBE conference 2020, EEA Virtual Congress 2020, Symposium for Social Cohesion in Ethnically Diverse Societies in Tilburg and ERINN Annual Conference 2021 for many helpful comments. The data collection was supported by a grant from the Czech Science Foundation (17-13869S), by the Slovak Research and Development Agency (APVV-0125-12 and APVV-19-0329), and by the Max Planck Institute for Tax Law and Public Finance. Bauer and Chytilová acknowledge support from ERC-CZ/AV-B (ERC300851901) and from the Czech Science Foundation (20-11091S). Cahlíková acknowledges support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866. The research was approved by the Ethical committee at the Technical University of Košice (12/7/2016).

[1] Ethnic minorities are targets of violence in many parts of the world (Horowitz, 1985; Yanagizawa-Drott, 2014). Historical evidence suggests that aggressive behaviours, including pogroms and attempted genocides, increase during periods of social and economic unrest within the majority group (Voigtlander and Voth, 2012; Anderson *et al.*, 2017; Grosfeld *et al.*, 2020).

This paper provides the first controlled experimental test of whether scapegoating behaviour is more common when people can displace punishment on members of a negatively stereotyped minority group than on members of their own group. Small-scale experiments in social psychology have shown that some people displace aggressive behaviour on innocent individuals if the provoking agent is unavailable (for a review, see Marcus-Newhall *et al.*, 2000).[2,3] We contribute by studying negative indirect reciprocity in an economic experiment implemented on a relatively large sample, and by using a rich experimental design that involves exogenous manipulation of the real-life identity of the scapegoat. We show that the identity of the scapegoat indeed matters: people's tendency to shift punishment is magnified if a scapegoat is from the Roma, an economically and socially disadvantaged minority group.

Cleanly identifying scapegoating behaviour with observational data is empirically challenging. First, it is nearly impossible to rule out the role of the standard economic incentives to harm innocent individuals, such as self-interested plundering of resources. In addition, in most real-life situations there is an element of uncertainty about who originated the harm. Members of the dominant group may punish innocent individuals from minority groups, because they (over)attribute responsibility for misfortunes to actions of minority groups. A controlled experimental environment allows eliminating these confounding factors. An experimental set-up allows a researcher to (i) measure how people behave when punishment of wrongdoers is inhibited, and people can punish only individuals who could not have causally contributed to the original harmful act, (ii) measure punishment responses in one-shot anonymous interactions that are costly for the punisher, and that provide no scope for material benefits of punishment, and (iii) compare behaviour towards a weaker minority group and towards the own (majority) group.

We therefore introduce a novel experimental paradigm, *punishing the scapegoat game*. In this game, impartial spectators can impose a monetary punishment on others at their own cost, after observing that someone malevolently destroyed the earned income of an individual from their own group. Incentivised experiments on punishment of socially undesirable behaviour focus exclusively on direct punishment of individuals who make active decisions whether or not to violate a social norm, including the third-party punishment game (Fehr and Fischbacher, 2004; Bernhard *et al.*, 2006). Instead, in the punishing the scapegoat game, we add a fourth person, the passive scapegoat who does not know the wrongdoer and who is not involved in any way in the original wrongdoing. This allows us to separate the person who commits a harmful act and a person whom the impartial spectator can punish. We inhibit the possibility to directly punish the wrongdoer, and are interested in how behaviour towards the scapegoat is affected by whether and how much harm a wrongdoer caused to a victim from the spectator's own ethnic group.

Importantly, we exogenously manipulate information about the ethnicity of the scapegoat. We are primarily interested in whether scapegoating behaviour of the dominant group is particularly strong when the target is a member of a negatively stereotyped, weaker group, a behavioural pattern we refer to as *minority scapegoating*. Existence of minority scapegoating implies that

[2] Social psychology experiments use different methods to expose subjects to frustrating situations, including derogatory comments from actors, putting their hand in cold water or working on a task in the presence of loud noise. Aggressive behaviour is typically measured by the willingness to apply an electric shock or noise blasted on a confederate (Baron and Bell, 1975; Marcus-Newhall *et al.*, 2000; Reidy *et al.*, 2010). In this paper, we design an economic experiment, in which interactions are anonymous, there is no deception, harming innocent persons is costly for the decision-maker, and both the provocation and the harm are pecuniary.

[3] Economic experiments have so far been designed to document the existence of *positive indirect* reciprocity (rewarding kind acts with kind acts towards other individuals) (Dufwenberg *et al.*, 2001; Engelmann and Fischbacher, 2009). We focus on studying *negative indirect* reciprocity (responding to hostile acts, by engaging in hostile behaviour towards unrelated parties).

we should observe a positive interaction effect on punishment of the scapegoat between the extent of harm done by the wrongdoer and the scapegoat being from the minority group. A noteworthy feature of using such a difference-in-difference approach is that we can test whether out-group bias in punishment happens above and beyond out-group bias in circumstances, when the dominant group does not respond to harm happening to their own group.

Eastern Slovakia is an apt natural setting to explore this phenomenon. The Roma people constitute the largest ethnic minority in Europe, estimated at 10–12 million persons. The average education levels of Roma are low (only 20% finish upper-secondary education). They are poorly integrated into labour markets (less than one-third are in paid employment), live in substandard housing and have lower life expectancy than the majority populations. It is estimated that 85% of Roma in Europe live below national poverty lines. Research shows that the Roma are subject to prejudice, and face discrimination in labour and housing markets (Bartoš *et al.*, 2016). In Eastern Slovakia, the Roma represent around 15% of the local population, and around 65% of them live segregated from the majority population.

We find that a non-negligible fraction of punishers (23%) from the dominant group shift punishment onto scapegoats (bystanders) when they cannot punish the wrongdoer. The main finding is that when wrongdoers harm the victim, the destructive behaviour towards the scapegoat doubles when the scapegoat is from the Roma minority than when the scapegoat is from the majority group. When wrongdoers do not harm the victim, ethnic majority punishers do not behave less favourably towards scapegoats from the ethnic minority as compared to the majority. Therefore, discrimination against the Roma minority by the dominant group *arises* only when wrongdoers harm the victim. Furthermore, among Roma decision-makers we do not find evidence of greater punishment of a scapegoat bystander from the majority group compared to Roma scapegoats. These results are in line with the interpretation that shifting of punishment onto innocent individuals is psychologically easier when the target is a member of a negatively stereotyped and weaker group (Bettelheim and Janowitz, 1950; Allport, 1954).

We consider several alternative mechanisms behind the main finding. We show that magnified harming of scapegoats from the minority group in response to injustice faced by members of the dominant group cannot be explained by collective punishment, i.e., situations in which the wrongdoer and the scapegoat both come from the Roma minority. Next, we argue that over-attribution of responsibility stemming from uncertainty about who committed the wrongdoing cannot explain our findings either, by virtue of the experimental design. Furthermore, the patterns we observe cannot also simply be an outcome of stable unconditional spite against the minority that would manifest itself under any circumstances. Our results are consistent with discriminatory preferences being latent in 'peaceful' times, but *activated* in environments when decision-makers respond to harm done to someone from their own group.

Our paper is related to economic experiments that study existence of discriminatory preferences based on ethnicity or socio-economic status (e.g., Fershtman and Gneezy, 2001; Falk and Zehnder, 2013; Bauer *et al.*, 2018; Berge *et al.*, 2020). We show that manifestations of discrimination against an economically disadvantaged, ethnic minority may depend on the decision environment, and be more pronounced when decisions happen in environments characterised by injustice happening to someone from the dominant group.[4] Furthermore, earlier work made

[4] A surprisingly large fraction of lab or lab-in-field experiments do not detect ethnic discrimination in standard experimental tasks (for a meta-study, see Lane, 2016), including experiments implemented in settings with a history of ethnic conflicts (Berge *et al.*, 2020). This observation is consistent with the pattern we find: ethnic biases can be latent in normal circumstances, in which people do not openly discriminate. Yet, such discriminatory preferences can be triggered

progress in studying out-group biases in the *direct* punishment of active norm violators (Bernhard *et al.*, 2006; Goette *et al.*, 2006; Schiller *et al.*, 2014). Most closely related to this paper, among indigenous tribes in Papua New Guinea, Bernhard *et al.* (2006) documented that third parties punish wrongdoers from an out-group more than wrongdoers from an in-group. In this paper we focus on biases in punishment of passive, innocent individuals in situations when punishment of wrongdoers is not possible, as is often the case in real life.

1. Experimental Design

1.1. *Sample*

The data collection took place in May–September 2017. The subjects of the majority Slovak ethnicity were sampled from the last two grades of the most common types of high schools (general, technical, business) in the Košice and Prešov regions, and at the campuses of the Technical University of Košice and the University of Prešov. The sample consists of 337 students aged 18–23. The sample characteristics are presented in Online Appendix Table A1. The subjects of the Roma minority ethnicity were sampled from 21 villages and towns in the region. The sample size is 484 young adults, aged 18–24. As expected, the Roma subjects had less education and came from a poorer socio-economic background (Online Appendix Table A2) compared to the subjects from the majority group.

1.2. *Experimental Tasks*

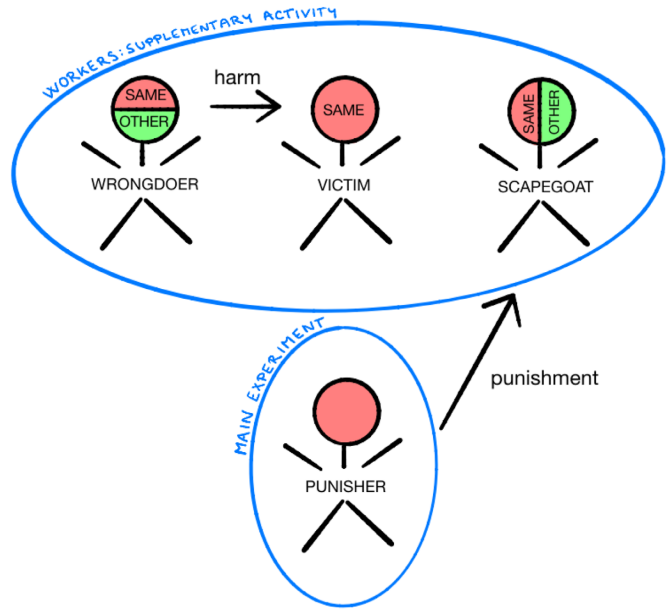
Figure 1(a) illustrates the *punishing the scapegoat game*. Each decision-maker (a punisher) is matched with three people—a wrongdoer, a victim and a scapegoat, neutrally labelled in the experiment as persons A, B and C—who come from different locations and do not know each other. The punisher is shown three pictures, each displaying 20 passport-style photos of people unknown to the punisher, homogeneous in terms of ethnicity, and taken against a neutral background. The punisher knows that s/he is matched with one person from each set of 20 photographs, but does not know with whom specifically.[5] The punisher is informed that each of these three people completed a work assignment and earned 8 euros for their work. Furthermore, the punisher learns that, after completing their work, the wrongdoer had an option to reduce the earnings of the victim by 0, 2, 4, 6 or 8 euros, and that the scapegoat was utterly passive.[6] We deliberately focus on punishment responses to a particularly malevolent behaviour: it reduces the earned income of the victim and does not create a pecuniary benefit for anyone, including the wrongdoer, and is thus designed to create a strong sense of injustice. Punishers further learn that only the wrongdoer had the option to reduce the earnings of someone else, and only the victim's earnings could have been reduced.

by situational factors that provide a scope for excuses or reduce self-control. This may suggest that standard economic experiments may underestimate the prevalence of discriminatory preferences.

[5] Displaying 20 individuals instead of one provides a sharp signal of ethnicity, retains anonymity since it remained unclear with whom specifically the decision-maker is matched, and allows us to avoid sympathies/antipathies towards a specific person in a picture driving decisions. Also, the individuals displayed in the pictures were homogeneous in terms of gender and age (all were young males, 18–23 years old), in order to avoid differential treatment based on these attributes.

[6] Before the experiment, we organised a supplementary work activity among a sample of different individuals, in order to make real the situation a punisher was confronted with, including the harm committed by the wrongdoer to the victim, and also to make the punisher's choices consequential. More information about the supplementary work activity is given in Online Appendix A.

(a) PUNISHING THE SCAPEGOAT GAME



(b) PUNISHING THE WRONGDOER GAME

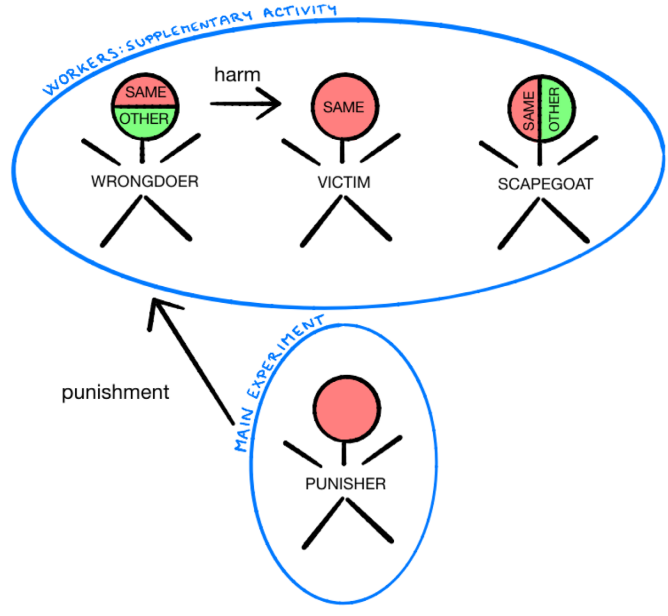


Fig. 1. *Illustration of the Experimental Tasks.*

Notes: The main sample are punishers, who come either from the majority group or from the Roma ethnic minority. Punishers learn about the harm committed by the wrongdoers towards the victims and can punish an innocent bystander—the scapegoat (panel (a))—or can directly punish the wrongdoer (panel (b)). The ethnic identities of the scapegoat and the wrongdoer are manipulated orthogonally: 'SAME' indicates that the player (wrongdoer/victim/scapegoat) is of the same ethnicity as the punisher, while 'OTHER' indicates that he is from the other ethnic group.

The task of the punisher is to decide whether and by how much to reduce the scapegoat's payment. Punishment is costly: reduction of each euro costs the punisher 0.10 euros. Punishers' decisions are elicited for all five possible actions of the wrongdoer towards the victim, using a strategy method. Importantly, the decision about the scapegoat's payoff was framed neutrally—the punishers were asked to decide whether to 'reduce the Person C's money by X euro by paying X∗10 cents'. In order to limit the scope for instrumental punishment of the scapegoat, the punishers were told that the wrongdoer would not be informed of the decision affecting the scapegoat. We set the punisher's endowment at 9 euros so that, even if s/he chose the maximum punishment level, his/her final payoff was 8.20 euros, i.e., higher than the payoff of all other players.

In addition, each punisher made a decision in the *punishing the wrongdoer game* (Figure 1(b)). The features of this task were identical to those of the *punishing the scapegoat game* (structure of payoffs, cost of punishment, visual design, strategy method), except that the punishers were asked to decide whether and by how much to reduce the wrongdoer's payoff. The two tasks were conducted in random order, and the participants did not know about the existence of the second task until after they finished the first one. Each task was payoff relevant at 10% probability.

1.3. *Manipulating Ethnic Identity*

We use photographs to signal ethnicity. Photographs provide a clear signal, because Roma people (who are of Indian origins and have a darker skin colour) are visually distinct from the Slovak majority. The victim is always of the same ethnicity as the punisher. In order to identify how ethnicity affects the decisions of the punishers, we exogenously manipulate signals of ethnicity of the scapegoat and of the wrongdoer. In the scapegoat SAME condition, the scapegoat has the same ethnicity as the punisher, whereas in the scapegoat OTHER condition, the scapegoat comes from the other ethnic group. Similarly, in the wrongdoer SAME condition, the wrongdoer is of the same ethnicity as the punisher, while in the wrongdoer OTHER condition, the wrongdoer comes from the other ethnic group. The signals of ethnicity of the scapegoat and of the wrongdoer are manipulated orthogonally. In this 2 × 2 'between-subject' design, each punisher is randomly allocated to one of the four possible combinations of wrongdoer SAME/OTHER and the scapegoat SAME/OTHER conditions. Randomisation checks indicate that the randomisation was successful (column 7 of Online Appendix Tables A1 and A2).

In the regression analyses, our main specification for estimating scapegoating is

$$\begin{aligned} Punishment\_Scapegoat_{ij} = {} & \beta_1 Harm\_intensity_{ij} + \beta_2 Scapegoat\_OTHER_{ij} \\ & + \beta_3 Harm\_intensity_{ij} \times Scapegoat\_OTHER_{ij} \\ & + \beta_4 Wrongdoer\_OTHER_{ij} + \gamma' \mathbf{X}_i + \varepsilon_{ij}, \end{aligned}$$

where i denotes the participant and j denotes the decision (each participant made five decisions).

Variable $Harm\_intensity_{ij}$ is the intensity of harm committed by the wrongdoer, $Scapegoat\_OTHER_{ij}$ and $Wrongdoer\_OTHER_{ij}$ are dummy variables, respectively indicating that the scapegoat and wrongdoer are from the other ethnic group. The main coefficient of interest is β_3, which for the decision-makers from the majority group reveals whether they engage in scapegoating behaviour more when the scapegoat is Roma (i.e., in minority scapegoating). Baseline controls $\mathbf{X}_i$ include gender and age of the punisher, and a dummy variable that the punishing

the wrongdoer game took place before the punishing the scapegoat game. Standard errors are clustered on the punisher level.

1.4. *Procedures*

We paid particular attention to maximise a correct understanding of the tasks. The experimenters explained the instructions one on one. The working/wrongdoing stage of the experiment and the decisions were explained in detail using a simple tablet interface (Online Appendix Figure A1). Before making decisions, the punisher had to answer six comprehension questions. If any of the answers were not correct, the experimenter explained the whole set-up and asked the comprehension questions once again. The level of understanding was high. On average, the decision-makers from the Slovak majority (Roma ethnic minority) answered 5.93 (4.27) comprehension questions correctly on the first attempt and 99.1% (76.9%) answered all comprehension questions correctly on the first or second attempt. We perform robustness checks with respect to comprehension in the analysis.

The subjects made their decisions anonymously and in private. Other participants could not hear the instructions or observe the decisions. Before each decision, the experimenter described the situation and gave the subject privacy to report their choices on the tablet computer.

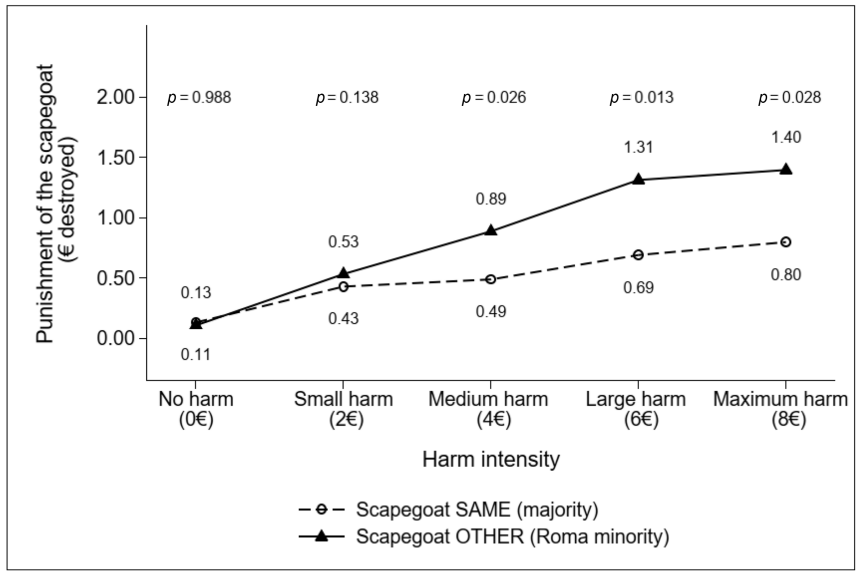
Online Appendix A provides more details about the sample and experimental design. Online Appendix Figure A2 shows the timeline of the data collection. The full experimental protocol is given in Online Appendix D.

2. Main Results

As we are interested in minority scapegoating, we first focus on punishers from the majority group. The main result of our paper is displayed in Figure 2, showing the punishment of the scapegoat across the five specific amounts of victim's earnings that the wrongdoer could decide to destroy (0, 2, 4, 6 and 8 euros). The dashed line shows the average amount of euros that the punishers decided to destroy in the SAME condition, when the scapegoat is also from the majority population, and the solid line shows the amount in the OTHER condition, when the scapegoat is a member of the Roma ethnic minority.

First, we see that the punishers are sensitive to the amount of harm done by the wrongdoer to the victim. The harm done to the scapegoat increases with the harm done by the wrongdoer. Second, (almost) no harm is done to the scapegoat when no harm is done by the wrongdoer. This is the case when subjects can harm a person from the majority Slovak group as well as from the Roma minority, suggesting that in 'peaceful' circumstances, people are not more inclined to harm the Roma, and thus people do not harbour unconditional spite towards Roma. Third, and perhaps most importantly, scapegoating against members of the minority is triggered when the punisher has observed harm done by the wrongdoer. In situations in which the wrongdoer harmed the victim, we find a systematic difference in responses between the scapegoat SAME and scapegoat OTHER conditions—punishment of scapegoats is twice as severe when the scapegoat is from the Roma minority than when the scapegoat is from the majority population. Specifically, in scapegoat SAME, an increase in harm intensity by one additional euro motivates punishers to lower the scapegoat's earnings by an additional 0.08 euros. In scapegoat OTHER, the effect doubles to 0.16 euros, and the difference between SAME and OTHER is statistically significant at the 1% level (p-value = 0.002, column (1) in panel A of Table 1). Because of such magnified punishment of

(a) INTENSITY OF PUNISHMENT



(b) PREVALENCE OF PUNISHMENT

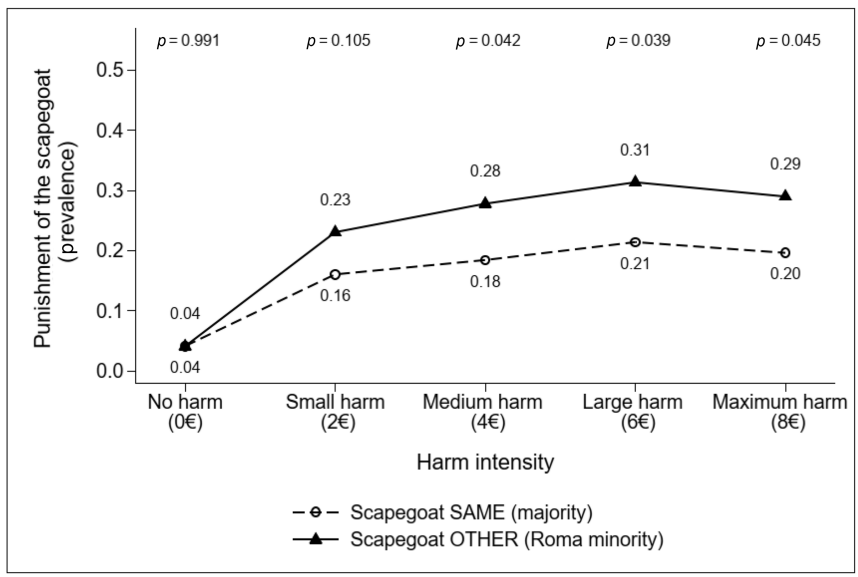


Fig. 2. *Punishment of the Scapegoat, by the Scapegoat's Ethnicity (Majority Sample).*
Notes: Mean punishment of the scapegoat (panel (a)) and the share of punishers who choose non-zero punishment of the scapegoat (panel (b)), by the ethnicity of the scapegoat and the harm caused by the wrongdoer to the victim. Punishers (and victims) are from the majority ethnic group. 'Scapegoat SAME' indicates that the scapegoat also comes from the majority ethnic group, while 'Scapegoat OTHER' indicates that scapegoat is ethnic Roma. Differences between the conditions are tested using the Wilcoxon rank-sum test in panel (a) and chi-squared test in panel (b); *p*-values are presented at the top. The sample is composed of punishers from the majority group.

Table 1. *Punishment of the Scapegoat.*

Dependent variable	Punishment of the scapegoat (intensity)	Punishment of the scapegoat (yes)	Punishment of the scapegoat (intensity)	Punishment of the scapegoat (intensity)	Punishment of the scapegoat (intensity)	Punishment of the scapegoat (intensity)
Sample	All	All	Punishment of the scapegoat = yes	All	All	All comprehension questions answered correctly
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: majority sample						
Harm intensity	0.08***	0.02***	0.20***	0.08***	0.08***	0.08***
	(0.02)	(0.00)	(0.07)	(0.02)	(0.02)	(0.02)
Scapegoat OTHER	−0.00	0.03	−0.57	−0.04	−0.06	−0.02
	(0.08)	(0.03)	(0.37)	(0.09)	(0.10)	(0.08)
Harm intensity × scapegoat OTHER	0.09***	0.01**	0.19**	0.09***	0.09***	0.08***
	(0.03)	(0.01)	(0.08)	(0.03)	(0.03)	(0.03)
Wrongdoer OTHER	−0.14	−0.03	−0.32	−0.14	−0.14	−0.16
	(0.13)	(0.03)	(0.24)	(0.13)	(0.13)	(0.13)
Controls	Baseline	Baseline	Baseline	Extended	Full	Baseline
Mean baseline (scapegoat SAME, 0 harm)	0.13	0.04	3.14	0.13	0.13	0.13
Observations	1,685	1,685	329	1,685	1,685	1,670
R^2	0.072	0.050	0.240	0.133	0.167	0.070
Panel B: Roma minority sample						
Harm intensity	0.09***	0.01***	0.16***	0.09***	0.09***	0.10***
	(0.02)	(0.00)	(0.04)	(0.02)	(0.02)	(0.02)
Scapegoat OTHER	−0.02	0.03	−0.38	0.01	0.03	−0.12
	(0.16)	(0.04)	(0.27)	(0.16)	(0.16)	(0.17)
Harm intensity × scapegoat OTHER	0.01	−0.00	0.03	0.01	0.01	0.01
	(0.03)	(0.00)	(0.05)	(0.03)	(0.03)	(0.03)
Wrongdoer OTHER	0.04	−0.02	0.27	−0.01	0.02	0.04
	(0.17)	(0.04)	(0.18)	(0.16)	(0.16)	(0.18)
Controls	Baseline	Baseline	Baseline	Extended	Full	Baseline
Mean baseline (scapegoat SAME, 0 harm)	0.97	0.26	3.77	0.97	0.98	0.80
Observations	2,420	2,420	906	2,420	2,410	1,860
R^2	0.017	0.008	0.064	0.093	0.121	0.021

Notes: OLS, standard errors (in parentheses) are clustered at the punisher level. *** $p < 0.01$ and ** $p < 0.05$. Punishers (and victims) in panel A are from the majority ethnic group and in panel B from the Roma minority. The dependent variable in columns (1), (3), (4)–(6) is the extent of punishment of the scapegoat (0–8 euros). In column (2), the dependent variable indicates that the punisher chose non-zero punishment of the scapegoat. 'Harm intensity' captures the harm caused by the wrongdoer to the victim (0–8 euros). 'Scapegoat OTHER' indicates that the scapegoat comes from a different ethnic group (Roma minority) from the punisher. 'Wrongdoer OTHER' indicates that the wrongdoer comes from a different ethnic group from the punisher. Baseline controls include the gender and age of the punisher, and a dummy variable indicating that the punishing the wrongdoer game took place before the punishing the scapegoat game. Extended controls also include experimenter fixed effects, a dummy variable indicating that the punisher is a university student (versus a secondary school student), location fixed effects, education of parents (dummy variables for mother/father with a university degree, dummy variables for education unknown) and a dummy variable indicating that the subject answered all control questions correctly on the first or second attempt. Full controls additionally include all other variables presented in Online Appendix Tables A.1 (majority sample) and A.2 (Roma minority sample). In column (6), we exclude all subjects who did not answer all control questions correctly at the first or second attempt. Online Appendix Table A5 shows the results of an estimation in which we use the triple interaction (harm intensity × scapegoat OTHER × majority sample) to study whether the bias against OTHER scapegoat is greater for the punishers from the majority group as compared to the Roma minority punishers (the coefficient is equal to 0.08 and p-value = 0.053 with baseline controls).

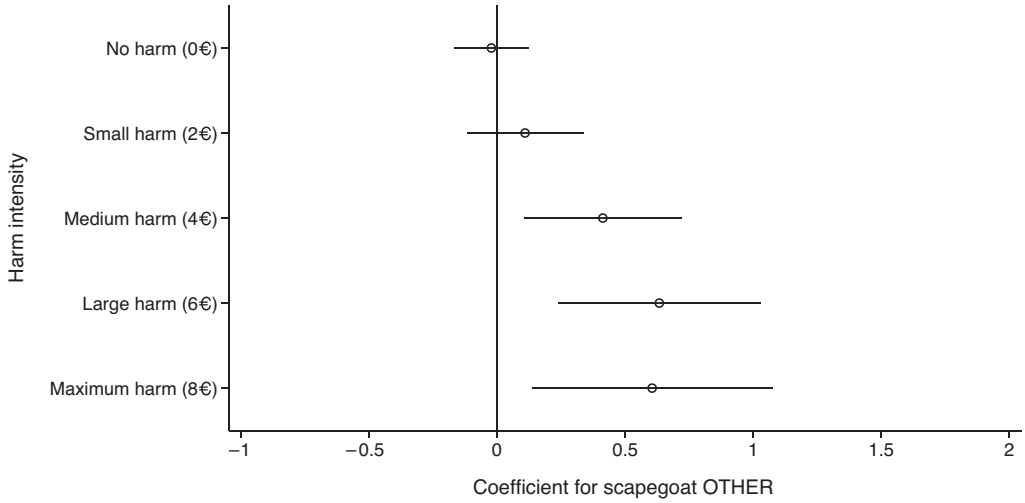


Fig. 3. *Discrimination against OTHER Scapegoats (Majority Sample).*
Notes: Estimated coefficients for 'Scapegoat OTHER', with 95% confidence intervals. The dependent variable is the extent of punishment of the scapegoat (0–8 euros). The coefficients are estimated by separate OLS regressions for the five possible levels of harm caused by the wrongdoer to the victim. We control for the gender and age of the punisher, and for a dummy variable indicating that the punishing the wrongdoer game took place before the punishing the scapegoat game. Punishers (and victims) are from the majority ethnic group. 'Scapegoat OTHER' indicates that the wrongdoer comes from a different ethnic group (Roma minority) from the punisher. The sample is composed of punishers from the majority group.

the scapegoat in OTHER, point estimates of discrimination against the ethnic minority gradually rise with greater harm intensity, and become statistically significant for situations when the wrongdoer destroyed 4, 6 or 8 euros of the victim's earnings (Figure 3, p-values = 0.008, 0.002, 0.011, respectively).

The interaction effect of the wrongdoer's level of harm committed and scapegoat OTHER on punishment of the scapegoat is driven by the extensive as well as the intensive margins (columns (2) and (3) in panel A of Table 1), and in both cases it is statistically significant at the 5% level. Each additional euro destroyed by the wrongdoer leads to an increase in the proportion of those who decide to punish the scapegoat, by 1.8 percentage points in scapegoat SAME and by 2.9 percentage points in scapegoat OTHER. Among those who decide to scapegoat, the amount destroyed from the scapegoat's earnings increases by 0.20 euros in scapegoat SAME and by 0.39 euros in scapegoat OTHER.

The results are robust to controlling for various additional variables. In Online Appendix Table A3, columns (1)–(9), the coefficients for harm intensity and for the interaction between harm intensity and scapegoat OTHER hardly change when we control for design features, including experimenter fixed effects, the subject's descriptive characteristics, the subject's understanding and location fixed effects. The results are similar if we use a non-linear specification for the harm caused by the Wrongdoer (columns (1)–(2) of Online Appendix Table A4).

Some of the literature suggests that scapegoating is more common when decision-makers from the dominant group interact with members of negatively stereotyped, vulnerable and smaller groups (Bettelheim and Janowitz, 1950; Allport, 1954), rather than vice versa. To test this idea, we

compare the behaviour of punishers from the majority group with the behaviour of punishers from the Roma minority. Indeed, we find weaker evidence of biases in punishments by the minority group against the majority group, as compared to biases of the majority group against the minority group. Specifically, Roma subjects punish the scapegoat, but we do not find evidence of co-ethnic bias (Online Appendix Figure A3)—the coefficient for the interaction between harm intensity and scapegoat OTHER is small in magnitude and not statistically significant (p-value = 0.737, column (1) of panel B of Table 1).[7] In Online Appendix Table A5, we use an estimate with triple interactions to show that the bias against OTHER scapegoat is greater for the punishers from the majority group as compared to the Roma minority punishers (p-value = 0.053 with baseline controls, column (1)).

The use of the strategy method to elicit choices for various possible actions of the wrongdoer has the advantage of providing a rich picture of the punisher's behaviour. Nevertheless, this approach may induce punishers towards greater differentiation in behaviour in different situations, and thus lead to greater observed sensitivity of punishers to the intensity of the harm done by the wrongdoer. Our main focus, however, is on estimating the *differences* in the punisher's sensitivity to harm intensity across the scapegoat's ethnicity (SAME versus OTHER). Since the SAME and OTHER conditions were implemented using a between-subject design and the decision environment was identical, except for the scapegoat's ethnicity, subjects could not be induced to differentiate between the punishment of scapegoats of the majority and minority ethnicities.[8] Thus, we believe that it is unlikely that any experimenter demand effect could explain the magnified punishment of the scapegoat in OTHER as compared to SAME.

Finally, we analyse choices in the *punishing the wrongdoer game*. We first discuss results for decision-makers from the Slovak majority group. Also in this task, the greater the harm caused by the wrongdoer, the stronger the punishment (Online Appendix Figure A4). The punishment response is approximately around five times stronger than in the *punishing the scapegoat game*. When no harm is done by the wrongdoer, we do not detect any discrimination against the Roma minority. The sensitivity to harm intensity is again systematically larger in wrongdoer OTHER as compared to SAME. The coefficients for an interaction term between harm intensity and wrongdoer OTHER are robust and statistically significant at the 1% level, both for the extensive and the intensive margins (panel A of Online Appendix Table A6 and Table A7). The magnified revenge in OTHER as compared to SAME gives rise to discrimination against the Roma minority, when the harm is large (4, 6 or 8 euros; Online Appendix Figure A5).

For punishers from the Roma minority, we find a somewhat weaker but qualitatively similar pattern as for the majority population (Online Appendix Figure A6). When no harm is done by the wrongdoer, we find no evidence of discrimination against the majority population, but harmful actions of the wrongdoer against a Roma subject trigger magnified revenge towards OTHER (majority) wrongdoers as compared to SAME (Roma) wrongdoers. The coefficient for an interaction term between the harm intensity and OTHER is statistically significant at the 10%

[7] It is reassuring that, when we exclude the 23% of subjects who demonstrated imperfect understanding, the estimates are similar to the original results (column (6) in panel B of Table 1 and Online Appendix Table A6).

[8] Many existing experiments studied differences in choices when the strategy method versus the direct-response method is used. Jordan *et al.* (2016) focused specifically on the third-party punishment game, and found that the use of the strategy method does not influence punishment decisions. Brandts and Charness (2011) provided an overview of 29 studies focusing on various experimental tasks. They found that, although the use of the strategy method affected the levels of behaviour in some of the tasks, in all of the experiments, a treatment effect identified with the strategy method was also observed with the direct-response method.

level (p-value = 0.097, column (1) in panel B of Online Appendix Table A6).[9] To sum up, when no harm is done to a member of the own group, the decision-makers do not discriminate against members of a different ethnicity. At the same time, wrongdoers of a different ethnicity are punished more severely than wrongdoers from the own ethnic group, for the same harmful actions.

3. Discussion: Mechanisms and Limitations

In this section we discuss several alternative explanations why shifting of punishment becomes larger when the scapegoat is Roma. We also acknowledge some limitations of our design and findings.

First, by virtue of the experimental design, differences in beliefs about who committed the harm or about future retaliation (statistical discrimination) are unlikely to explain our findings. Punishers faced no uncertainty about who was responsible for the wrongdoing, since the experimental protocol and graphical aids made it clear that wrongdoers caused the harm to the victim, while scapegoats did not (Online Appendix Figure A1). Thus, minority scapegoating is unlikely to arise because of over-attribution of responsibility for wrongdoing to Roma. Next, the observed bias in punishment is unlikely to be driven by differences in beliefs about future interactions, perhaps out of fear of facing greater likelihood of revenge from in-group members. Punishers knew scapegoats and wrongdoers would not have any opportunity to take revenge after their punishment decision, and the interactions were one shot and anonymous, with scapegoats and wrongdoers coming from different locations.

Second, the results do not support the interpretation that the observed magnified punishment of scapegoats from the Roma minority could be due to the notion of collective responsibility, which refers to retaliation directed, not only against the wrongdoer, but also against other members of his group who have no direct association with the perpetrator (Lickel *et al.*, 2003; Cushman *et al.*, 2012). In our experiment, collective responsibility would predict that greater punishment of minority scapegoats is triggered only in a situation where the wrongdoer is also from the minority group, but not when the wrongdoer is from the majority group. To test this, we take advantage of the orthogonal experimental variation of identity of the wrongdoer and of the scapegoat. We find that the minority scapegoats are more harshly punished regardless of the ethnic identity of the wrongdoer (Online Appendix Figure A7 and Table A9). In a regression analysis, we restrict the sample to subjects who were informed that the wrongdoer was from the majority group and still find an ethnic bias in punishment of the scapegoat. In other words, the minority scapegoats are more prone to face harm for injustice done to the member of the majority group, even when this injustice originates from within the majority group.

Third, plain unconditional spite towards Roma cannot fully explain our findings either, because it would imply that decision-makers should treat the ethnic minority systematically more harshly than members of their own group, independently of the social context. This is not the case, since

[9] We cannot rule out that the out-group bias in the punishment of the wrongdoer is the same for punishers from the Slovak majority and the Roma minority groups (Online Appendix Table A8, coefficient for the interaction (harm intensity × wrongdoer OTHER × majority sample)). While the role of identity is qualitatively similar, the strength of the effects and the extent of punishment differ across the decision-makers' ethnicity. The extent of punishment is higher among Roma than among ethnic majority decision-makers when no harm is committed by the wrongdoer (0.91 versus 0.24 euros, respectively), and lower in the situation of maximum harm (2.71 versus 5.23 euros, respectively). Thus, Roma decision-makers were less sensitive to harm intensity.

we find no evidence of discrimination against the minority, as long as there is no harm committed against a victim.

While we can rule out several alternative explanations, we cannot empirically differentiate the following psychological mechanisms. First, if deep spiteful urges to harm the Roma are controlled at normal emotional states, witnessing blatant injustice may trigger anger and reduce such self-control, leading to more anti-social behaviour. A natural next step in testing this psychological channel would be to elicit punishment decisions and at the same time measure emotions. Second, observing unethical behaviour of the wrongdoer can reduce the costs to the self-image of the punishers from acting based on latent discriminatory preferences, because harmful behaviour becomes perceived as more acceptable. Finally, it is also possible that reduced earnings of the victim from the dominant group below the level obtained by the scapegoat provides a rationale or an excuse for destroying income of a member of the Roma ethnic minority, due to its relatively lower socio-economic status. Specifically, given that outside of the experiment Roma people are more likely to be unemployed and receive social benefits than members of the majority group, decision-makers from the majority group may perceive it as unfair if Roma receive greater earnings for the same task, as compared to workers from a majority group.

We note several potential limitations of our research design and fruitful areas for future research. First, while there are clear advantages of identifying discriminatory behaviour towards an important, real-life minority group like the Roma people, this approach comes at a cost of lacking control over the dimension based on which people may discriminate: ethnicity, culture or socio-economic status. Although we have carried out several steps that aim to disentangle the effect of ethnicity from socio-economic status (see Online Appendix B), this aspect remains an open question. Since ethnicity-based discrimination is often justified based on socio-economic arguments, a clean test of whether greater displacement of punishment can arise purely based on socio-economic status, would require exogenous variation in information about the socio-economic status of the scapegoat, while holding ethnicity the same. In general, more research is needed to explore which specific group attributes make certain individuals or groups convenient scapegoats, and also to assess whether the minority scapegoating observed against the Roma in Eastern Slovakia is generalisable to other settings.

Second, all experimenters came from the Slovak majority group. Even though the design prevented them from observing subjects' decisions, the ethnic identity of the experimenter may have perhaps influenced the behavioural responses. An interesting question is whether the patterns, especially behaviour of the Roma minority, are robust to having the protocol implemented by Roma experimenters.

Finally, we deliberately created a decision situation, in which punishers face a clear injustice that creates a strong urge to punish. We elicited responses to observing unambiguously nasty behaviour—malevolent destruction of the victim's earned income—whereas the standard third-party punishment game (TPP) measures punishment responses to observing a lack of willingness to share, a much weaker form of norm violation. Furthermore, punishment in our experiment is relatively cheap (0.1 euros reduce earnings by 1 euro), while in the standard TPP game the ratio is 1:3. Consequently, we observe relatively high punishment levels of wrongdoers, as compared to existing TPP (Fehr and Fischbacher, 2004; Kosfeld and Rustagi, 2015). Thus, a natural question is whether we would arrive at qualitatively similar patterns if we used higher costs of punishing others or different forms of injustice faced by the dominant group.[10]

[10] We find it reassuring that we replicate greater punishment of out-group as compared to in-group norm violators identified in earlier work (Bernhard *et al.*, 2006; Schiller *et al.*, 2014)—see Online Appendix Table A6—suggesting that

4. Conclusion

Scapegoating has been considered by social scientists to be an important mechanism in the emergence of pogroms, witch-hunts and large-scale violence against unpopular and weaker minority groups. This paper provides the first controlled, experimental test of how scapegoating behaviour is shaped by the group identity of the scapegoat. We use a new incentivised task, *punishing the scapegoat game*, implemented in East Slovakia to uncover how observing harmful actions against members of one's own group shapes the punishment of innocent individuals. We show that the identity of the scapegoat indeed matters: punishers from the ethnic majority group systematically punish innocent members of the Roma minority group for harmful actions committed by other people.

Our findings have several potentially important implications. First, we show that pure observation of injustice and wrongdoing against the individual's own group activates latent discriminatory preferences, both when treating innocent individuals as well as wrongdoers. This indicates that courts, and other settings in which people make punishment choices, are particularly discrimination-prone environments, in line with evidence of strong biases against minorities in judicial decisions (Shayo and Zussman, 2011; Alesina and La Ferrara, 2014; Rehavi and Starr, 2014). Second, the results suggest that ethnic minorities are at greater risk of facing aggressive behaviour when social problems within the dominant group become salient features of the environment. So far, economists have typically attributed sudden spikes in aggressive behaviour towards weaker groups to changes in economic incentives or beliefs about the likelihood of facing a penalty for aggressive behaviour (Miguel 2005; Blattman and Miguel, 2010; Grosfeld *et al.*, 2020), assuming that revealed (anti-)social preferences towards other groups are stable. In our experiment, economic incentives are held constant and thus cannot explain the scapegoating behaviour observed. Of course, this does not imply that economic incentives do not play an important role in real-life aggression towards minority groups. However, our evidence suggests that that may not be the complete picture. It strengthens the case for taking seriously psychological channels, including scapegoating, through which deterioration of the social environment may fuel inter-group conflicts.

CERGE-EI (a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences) and Institute of Economic Studies, Faculty of Social Sciences, Charles University, Czech Republic
University of Bonn, Germany
Institute of Economic Studies, Faculty of Social Sciences, Charles University and Economics Institute of the Czech Academy of Sciences, Czech Republic
UC Berkeley, CEPR & NBER, USA
Technical University of Košice, Slovak Republic & Durham University, UK

Additional Supporting Information may be found in the online version of this article:

Online Appendix
Replication Package

although the differences in parameters of TPP affect the level of punishment, they are unlikely to shape the qualitative results on the role of group identity of the punished person.

References

Alesina, A. and La Ferrara, E. (2014). 'A test of racial bias in capital sentencing', *American Economic Review*, vol. 104(11), pp. 3397–433.

Allport, G.W. (1954). *The Nature of Prejudice*, New York: Addison-Wesley.

Anderson, R.W., Johnson, N.D. and Koyama, M. (2017). 'Jewish persecutions and weather shocks: 1100–1800', Economic Journal, vol. 127(602), pp. 924–58.

Baron, R.A. and Bell, P.A. (1975). 'Aggression and heat: Mediating effects of prior provocation and exposure to an aggressive model', *Journal of Personality and Social Psychology*, vol. 31(5), pp. 825–32.

Bartoš, V., Bauer, M., Chytilová, J. and Matějka, F. (2016). 'Attention discrimination: Theory and field experiments with monitoring information acquisition', *American Economic Review*, vol. 106(6), pp. 1437–75.

Bauer, M., Cahlíková, J., Chytilová, J. and Želinský, T. (2018). 'Social contagion of ethnic hostility', *Proceedings of the National Academy of Sciences*, vol. 115(19), pp. 4881–6.

Berge, L.I.O., Bjorvatn, K., Galle, S., Miguel, E., Posner, D.N., Tungodden, B. and Zhang, K. (2020). 'Ethnically biased? Experimental evidence from Kenya', *Journal of the European Economic Association*, vol. 18(1), pp. 134–64.

Bernhard, H., Fischbacher, U. and Fehr, E. (2006). 'Parochial altruism in humans', *Nature*, vol. 442(7105), pp. 912–15.

Bettelheim, B. and Janowitz, M. (1950). *Dynamics of Prejudice: A Sociological Study of Veterans*, New York: Harper and Brothers.

Blattman, C. and Miguel, E. (2010). 'Civil war', *Journal of Economic Literature*, vol. 48(1), pp. 3–57.

Brandts, J. and Charness, G. (2011). 'The strategy versus the direct-response method: A first survey of experimental comparisons', *Experimental Economics*, vol. 14(3), pp. 375–98.

Cushman, F., Durwin, A.J. and Lively, C. (2012). 'Revenge without responsibility? Judgments about collective punishment in baseball', *Journal of Experimental Social Psychology*, vol. 48(5), pp. 1106–10.

Doob, L.W., Miller, N., Mowrer, O.H., Sears, R. and Dollard, J. (1939). *Frustration and Aggression*, New Haven, CT: Yale University Press.

Dufwenberg, M., Gneeyz, U., Güth, W. and van Damme, E. (2001). 'Direct vs indirect reciprocity: An experiment', *Homo Oeconomicus*, vol. 18, pp. 19–30.

Engelmann, D. and Fischbacher, U. (2009). 'Indirect reciprocity and strategic reputation building in an experimental helping game', *Games and Economic Behavior*, vol. 67(2), pp. 399–407.

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

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

Fershtman, C. and Gneezy, U. (2001). 'Discrimination in a segmented society: An experimental approach', *Quarterly Journal of Economics*, vol. 116(1), pp. 351–77.

Goette, L., Huffman, D. and Meier, S. (2006). 'The impact of group membership on cooperation and norm enforcement', *American Economic Review*, vol. 96(2), pp. 212–16.

Grosfeld, I., Sakalli, S.O. and Zhuravskaya, E. (2020). 'Middleman minorities and ethnic violence: Anti-Jewish pogroms in the Russian empire', *Review of Economic Studies*, vol. 87(1), pp. 289–342.

Horowitz, D.L. (1985). *Ethnic Groups in Conflict*, London: University of California Press.

Jordan, J., McAuliffe, K. and Rand, D. (2016). 'The effects of endowment size and strategy method on third party punishment', *Experimental Economics*, vol. 19(4), pp. 741–63.

Kant, I. (1965). *The Metaphysical Elements of Justice*, New York: Bobbs-Merill Company.

Kosfeld, M. and Rustagi, D. (2015). 'Leader punishment and cooperation in groups: Experimental field evidence from commons management in Ethiopia', *American Economic Review*, vol. 105(2), pp. 747–83.

Lane, T. (2016). 'Discrimination in the laboratory: A meta-analysis of economics experiments', *European Economic Review*, vol. 90, pp. 375–402.

Lickel, B., Schmader, T. and Hamilton, D.L. (2003). 'A case of collective responsibility: Who else was to blame for the Columbine high school shootings?', *Personality and Social Psychology Bulletin*, vol. 29(2), pp. 194–204.

Marcus-Newhall, A., Pedersen, W.C. and Carlson, M. (2000). 'Displaced aggression is alive and well: A meta-analytic review', *Journal of Personality and Social Psychology*, vol. 78(4), pp. 670–89.

Miguel, E. (2005). 'Poverty and witch killing', *Review of Economic Studies*, vol. 72(4), pp. 1153–72.

Rehavi, M.M. and Starr, S.B. (2014). 'Racial disparity in federal criminal charging and its sentencing consequences', *Journal of Political Economy*, vol. 122(6), pp. 1320–54.

Reidy, D.E., Foster, J.D. and Zeichner, A. (2010). 'Narcissism and unprovoked aggression', *Aggresisive Behavior*, vol. 36(6), pp. 414–22.

Schiller, B., Baumgartner, T. and Knoch, D. (2014). 'Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination', *Evolution and Human Behavior*, vol. 35(3), pp. 169–75.

Shayo, M. and Zussman, A. (2011). 'Judicial ingroup bias in the shadow of terrorism', *Quarterly Journal of Economics*, vol. 126(3), pp. 1447–84.

Voigtlander, N. and Voth, H.-J. (2012). 'Persecution perpetuated: The medieval origins of anti-semitic violence in Nazi Germany', *The Quarterly Journal of Economics*, vol. 127(3), pp. 1339–92.

Yanagizawa-Drott, D. (2014). 'Propaganda and conflict: Evidence from the Rwandan genocide', *Quarterly Journal of Economics*, vol. 129(4), pp. 1947–94.