A Collective Turn in the Philosophy of Hate Speech

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Philosophie an der
Ludwig-Maximilians-Universität München
in Co-tutelle mit der
Universidad de Granada

vorgelegt von

Jimena Zapata

2024

Reviewers

From Ludwig–Maximilians–Universität München (Germany)

First Reviewer:
Prof. Dr. Ophelia Deroy
Chair for Philosophy of Mind
Ludwig–Maximilians–Universität München
Faculty of Philosophy, Philosophy of Science and the Study of Religion

Second Reviewer:
Prof. Dr. Monika Betzler
Chair for Practical Philosophy and Ethics
Ludwig–Maximilians–Universität München
Faculty of Philosophy, Philosophy of Science and the Study of Religion

From University of Granada (Spain)

First Reviewer:
Dr. Alberto Neftalí Villanueva Fernández
Department of Philosophy I
University of Granada

Second Reviewer
Dr. Ivar Allan Rodríguez Hannikainen
Department of Philosophy I
University of Granada

Date of submission: October, 11th 2023
Date of oral exam: December, 8th 2023

A Collective Turn in the Philosophy of Hate Speech

Tesis presentada por

Jimena Zapata

Para optar al grado de doctora con mención internacional en el Programa de Doctorado en Filosofía (BO2.56.1).

Directores: Prof. Dr. Ophelia Deroy (LMU)
Dr. Neftalí Villanueva Fernández (UGR)

UNIVERSIDAD
DE GRANADA

Facultad de Filosofía y Letras
Departamento de Filosofía I

2024

XIII
Mass

At the end of the battle the fighter lay dead. A man came to him
and said: 'Don't die! I love you too much!'
But the corpse, alas, went on dying.

Two came to him and again said:
'Don't leave us! Take heart!
Come back to life!'
But the corpse, alas, went on dying.

Then twenty, a hundred, a thousand,
Five hundred thousand, came, crying:
'So much love and yet so powerless against death!'
But the corpse, alas, went on dying.

Millions surrounded him,
pleading together:
'Brother, don't leave us!'
But the corpse, alas, went on dying.

Then, all the men on earth
stood round him. The corpse eyed them sadly,
overwhelmed. He got up slowly,
embraced the first man, started to walk…

César Vallejo
España, aparta de mí este cáliz, 1939.
Translation by Paul O'Prey

Acknowledgments

Thanks to all of you who inspired this adventure called PhD, whose curiously I got to know in collective frameworks: organising a demonstration and dreaming around a table back in 2011, hiking together the German-Austrian border, discussing books in a club, travelling to unknown places, tasting wines, starting a family, dancing and preparing an improvised "shelter" during a pandemic to continue researching, questioning assumptions, making predictions, imagining solutions and sharing a good brunch. You all made me realise that the best solutions to embedded problems are those that we perform together, not because they are always the most brilliant or efficient, but because they are inclusive and create the feeling of community and belonging in us. Thanks to my 15M team, especially Irene, Antonio and Mari Jose. Thanks to my group of MIS hikers for showing me that there is no mountain high enough if I can count on a team.

Thanks to the great UGR group, especially Neftalí Villanueva, my co-supervisor and friend, who once considered a good idea, I decided to start a PhD in philosophy. Thanks for all your feedback and suggestions, for listening to me for hours and for giving a chance to my ideas and projects. Also, thanks to Manu, Ivar, Maria José and Manolo for the feedback and brainstorming and for encouraging me to persist in this adventure. Thanks to all my colleagues at CVBE & Crowd-Cognition Labs, especially to Lenka, Louis, Lucas, Oriane, Dardo, Julia, Jurgis and Rebecca for being such supportive colleagues, always up for brainstorming with a coffee or a glass of wine. Many thanks to Bahador for transforming my ideas into drawings and helping me to see them better. More especially to Justin, thanks for your patience, for listening, for always finding time for my questions, for editing my writings, and for always being up for a glass of wine when I needed it the most. It is a gift calling you my friend and co-author.

Thanks to my research assistants, Cuici and Clemens, I cannot imagine how I could have finished this work if you two had not been working and learning by doing at my side. Thanks to everyone who helped with the revision, edition and formatting: Etienne, Chris, Carl, Julianna.

Special thanks to Rae Langton, whose generosity I will never forget. Thanks, Rae, for patiently supervising my efforts to understand Speech Act theory and for receiving me in beautiful Newnham College (University of Cambridge), letting me talk about my writings in the same room where, ninety-five years ago, Virginia Woolf presented hers. It was an honour. Thanks to the Mentoring Program at LMU, especially my mentor, Monika Betzler, for financially supporting all my posters and presentations abroad. Thanks to the Institute of Philosophy (University of London) for giving me the perfect writing place to finish this work.

Ophelia, my supervisor in Munich, deserves a separate paragraph. Thanks for inspiring us, Ophelia. Your outstanding ability to create a research team where every idea counts and everyone has access to space, knowledge, recommendations, inspiration, and fun amazes me. Thanks for letting me be one more in such a team. Thanks for checking every word of my writings, for each article sent at 8 am on a Saturday, for all the books, the travelling, the yoga, the great wine bottles and ginger shots. Thanks for finding time for feedback even at the airport before taking a plane and encouraging me to submit my proposals to workshops enthusiastically and confidently. Thanks for teaching me to celebrate every triumph, whether small or big. Thanks for helping me to carry out this project, which began just some months before a global pandemic. Thanks for being a mentor and a friend, but more importantly, for becoming a partner in Terra Cognita to continue researching, sharing knowledge and envisioning collective solutions.

Finally, thanks to my dad for sharing ideals and readings and Christian for always imagining me writing and reading by the sea and helping me to make a dream come true. Thanks to Leon, this work is dedicated to you to help you understand the importance of finding your group and making your dreams come true. Thank you, my little one, for sharing our time with my projects and ideas and patiently understanding my writing weekends, but also for making me stop and watch a film, go swimming or eat chocolate together. Thanks to my extended family, Coco, Rox, Teresa, Josefina and Pepe, for never letting me alone despite the physical distance.

Contents

Acknowledgments……………………………………………………………………..……I

Contents……………………………………………………………………………………...V

List of Figures……………………………………………………………………..…..…IX

Summary………………………………………………………………………………...XIII

Zusammenfassung…………………………………………………………………...XXI

Chapter 1. Characterising hate speech and its audience ……………..1

1.1. Doing harm with words……………………………………………………………..1

1.2. Theoretical framework……………………………………………………………...2

1.3. The audience of hate speech……………………………………………………...7

1.4. Methodological approach to hate speech harm…………………………….9

Chapter 2. Perceiving speech harm: The illusion of ordinary people's leniency against hate speech……………………..15

2.1. Introduction…………………………………………………………………………...15

2.2. Experimental work: Methods……………………………………………………25

2.2.1. Ethics Information……………………………………………………..25

2.2.2. Study Description……………………………………………………...25

2.2.3. Testing Materials………………………………………………………27

2.2.3.1. Experimental Vignettes…………………………………...27

2.2.3.2. Distractors………………………………………………………28

2.2.3.3. Attention task...29

2.2.4. Procedure..29

2.2.5. Pre-registered sampling plan, power analysis
and exclusion criteria...30

2.2.6. Pre-registered analysis and results....................................31

2.2.7. Exploratory analysis and results.......................................41

2.2.8. Discussion...42

2.2.9. Differences with pilot study results...................................45

2.2.10. Data availability..47

2.2.11. Code availability..47

2.2.12. Supplementary materials...48

2.3. Conclusion...56

Chapter 3. Assessing bystander’s responses to hate speech: Collective opposition is more effective than individual confrontation.......................................59

3.1. Introduction...59

3.2. Studying responses to hate speech through visual vignettes............61

3.3. Experimental work..64

3.3.1. Experiment 1..64

3.3.1.1. Study Description...64

3.3.1.2. Participants...66

3.3.1.3. Procedure...67

3.3.1.4. Testing Materials
(Visual Vignettes)..68

3.3.1.5. Attention-check Task...69

3.3.1.6. Analysis Strategy..70

3.3.1.7. Results...70

3.3.1.7.1. The incident's level of harm..................70

3.3.1.7.2. The perpetrators' deserved blame.....73
3.3.1.8. Exploratory analysis.....73
3.3.1.9. Discussion.....75
3.3.1.10. Supplementary materials.....76
3.3.2. Experiment 2.....84
3.3.2.1. Study Description.....84
3.3.2.2. Participants.....87
3.3.2.3. Testing Materials.....88
3.3.2.4. Attention-check Task.....90
3.3.2.5. Procedure.....90
3.3.2.6. Analysis Strategy.....91
3.3.2.7. Results.....93
3.3.2.7.1. The specific contribution of the target bystander to that harm.....93
3.3.2.7.2. The overall level of harm of the incident.....96
3.3.2.8. Discussion.....97
3.3.2.9. Supplementary materials.....98
3.3.3. General discussion.....102
3.3.4. Limitations to Generality.....103
3.3.5. Data & code availability.....104
3.3.6. Ethical approval.....104
3.3.7. Informed consent.....104
3.4. Conclusion.....105

Chapter 4. Revisiting hate speech harm to characterise the speaker.....107

4.1. Revisiting hate speech harm.....107

4.2. A lone wolf or a group member?...111

4.3. Group speakers and the right to freedom of speech…………………..114

4.4. A collective response to a group speech……………………………………118

Chapter 5. Conclusions…………………………………………………………..127

References……………………………………………………………………………...133

List of figures

Chapter 1

Figure 1: The audience of hate speech.

Chapter 2

Figure 1: Box plots show the appropriate punishment for each action type (N=1291).

Figure 2: Box plots show the appropriate punishment for each consequence type (N=1291).

Figure 3: Box plots show the likelihood of denouncing perpetrators for each action type (N=1291).

Figure 4: Box plots show the likelihood of denouncing perpetrators for each consequence type (N=1291).

Figure 5: Box plots show the level of harm for each action type (N=1291).

Figure 6: Box plots show the level of harm for each consequence type (N=1291).

Figure 7: Interaction plot between action type and consequence type in the level of harm inflicted on the victim (N=1291)

Chapter 3

Figure 1. Example visual vignettes for each of the 4 experimental conditions (Scenarios A-D). The perpetrator is labelled as “A”.

Figure 2. The attention check vignette (Experiment 1).

Figure 3. (a) A stacked bar chart showing the distribution of ratings for the incident’s level of harm, grouped by scenario; (b) grey bars show mean rating (and whiskers show 95% bootstrapped CIs) for the incident’s level of harm; blue diamonds and lines show median responses; (c) A stacked bar chart showing the distribution of ratings for the blame assigned to perpetrators, grouped by scenario; and (d) grey bars show mean rating (and whiskers show 95% bootstrapped CIs) for the blame assigned to perpetrators, blue diamonds and lines show median responses.

Figure 4. Graph showing the perpetrator supporters identified grouped by scenarios.

Figure 5. The image shows example visual vignettes for each of the 4 experimental conditions: Scenario A with 0 opposers, Scenario B with one, Scenario C with 2 and Scenario D with 3 opposers.

Figure 6. The image illustrates the target bystander in Scenario D. Such an image was presented alongside all questions about a bystander's individual contribution to the overall harm caused by the incident to ensure that participants knew which bystander was the focus of each question.

Figure 7. Responses are grouped by scenario, and the target bystander in each scenario is indicated with an arrow. (a) Stacked bar chart showing the rating distribution for the target bystander's contribution to harm (positive ratings = increase harm, negative ratings = reduce harm, zero = makes no difference). (b) Grey bars show the mean rating (and whiskers show 95% bootstrapped CIs) of the target bystander's contribution (increase or reduce) to the damage caused by the incident; blue diamonds and lines show median responses. (c) Stacked bar chart showing the rating distribution for the incident's overall level of perceived harm. (d) Grey bars show the mean rating (and whiskers show 95% bootstrapped CIs) of the incident's overall level of perceived harm; blue diamonds and lines show median response.

Summary

The present dissertation is divided into five Chapters. As an introduction, Chapter 1 characterises hate speech, the harm it creates and its audience. Throughout our investigation, we defend the idea that hate speech is a harmful mechanism used in intergroup disputes for social dominance (Charles-Toussaint & Crowson, 2010; Duckitt & Sibley, 2017; Hoover et al., 2021). Moreover, it targets people based on their actual or perceived "race", colour, descent, national or ethnic origin, age, disability, language, religion, sex, gender, sexual orientation and other identity features (Mihajlova et al., 2013). Therefore, its proliferation threatens coexistence within diverse societies, where people from various identities and backgrounds live together.

Following Speech Act theory, which defines "illocutionary" speech acts as those that do something by saying something (e.g., marrying someone by saying "Yes, I do"), we feature hate speech as such an act (Langton, 2018a; Maitra, 2012), defending the idea that its power to harm others is directly linked with its capacity to perform various harmful actions through speech.

By saying, "We do not want your kind here!", "Go back home!" or "Stop Islamization of our country!" along with other paradigmatic hate speech, we perform several actions: We rank minorities or disfavoured groups as inferior (Langton, 2012; Langton, 2018a; Langton, 2018b; Maitra, 2012), direct bystanders to side with hate speakers, order people to leave if they do not conform with the speaker's standard of a good citizen (Fraser, 2023; Lepoutre 2021; Waldron, 2012), encourage like-

minded fellows to take action against targeted groups and, sometimes, perform several or all those actions at a time (Lewiński, 2021).

Importantly, we defend the idea that most hate speech not only *causes* harm but *constitutes* harm to its targets (Langton, 2018a): whether ranking a group as inferior, blaming minorities for circumstances outside of their control, directing disfavoured groups to leave or legitimating particular treatment of them through hate speech, these all hurt people (Langton, 2018a). In addition, those actions can *cause* feelings of humiliation, helplessness, isolation, low self-esteem or anger (Fattoracci & King, 2023; King et al., 2011), harming people psychologically and physiologically to varying degrees (Eisenberger, 2015).

Moreover, we characterise hate speech as highly context-dependent (Moreno & Pérez Navarro, 2021). This quality contributes to it being perceived as directed at different subjects, expanding its audience and the people it harms, going beyond its direct targets to also include bystanders and society at large. Depending on the context, the audience of hate speech may be any of us. Notwithstanding, as its audience, we all play a relevant role in determining, modulating and countering the actions a hate speech act can perform. Therefore, a good starting point to address this phenomenon should count on all of us, as potential audience, to identify whether and to what extent we perceive hate speech as harmful and the conditions under which a response contributes to reducing its harm.

Following an experimental moral philosophy approach to substantiate our characterisation for hate speech and its harm, we conducted two studies testing ordinary people's intuitions in those regards, reproduced in Chapters 2 and 3.

Chapter 2 reproduces a Registered Report published in May 2023 in Scientific Reports (Springer Nature), co-authored with Prof. Dr. Ophelia Deroy. In this study, we tested whether people are intrinsically averse to verbal harm compared to other kinds of hate actions (nonverbal, bodily actions) and to what extent. Considering that bystanders rarely report hate speech incidents and the legal, theoretical and social hesitancy to punish them, we hypothesised that people would be more lenient against hate speech than nonverbal hate actions, which share intentions and consequences. We conducted an experiment with 1309 British citizens who read descriptions of verbal and nonverbal incidents stemming from identical hateful intent, which created the same consequences. We asked them how much punishment the speaker (perpetrator) should receive, how likely they would be to denounce such an incident and how harmful the actions were.

The results contradicted our pre-registered hypotheses and the predictions of dual moral theories, which hold that intention and harmful consequences are the sole psychological determinants of punishment. Instead, participants consistently rated verbal hate incidents as more deserving of punishment and denunciation, and as being more damaging than nonverbal incidents. The difference remained even when we shifted the scenarios to let participants know that the targets suffered no negative consequences (e.g., the target was deaf and so could not hear the racist remark).

We explain this difference through the concept of action aversion (Miller et al., 2014) in opposition to outcome aversion, suggesting that lay observers perceive something inherent to speech that makes them assess it as more harmful and deserving of punishment and denunciation as opposed to other nonverbal hate actions, regardless of its consequences. Later, in Chapter 4, we interpret the intrinsic feature captured by folk intuitions in hate speech as the ability to harm by saying (i.e., *constituting* harm, and not only *causing* it) and to target different people simultaneously (e.g., direct-targets and random bystanders), which remains even when the speakers do not manage to harm their direct targets.

Chapter 3 reproduces a paper submitted to the scientific journal *Humanities and Social Sciences Communications*, currently under review. Co-authored with Prof. Dr. Ophelia Deroy, Dr. Justin Sulik and Mr. Clemens von Wulffen, it explores people's perception of the role played by silent or opposing bystanders in reducing the harm caused by hate speech. We depart from two widespread assumptions: a prevailing passive attitude towards hate speech and the consideration of bystanders' opposition when facing a hate incident as helpful in mitigating its harm, something that has been stressed by governmental authorities, sociologists, and philosophers (Ayala & Vasilyeva, 2016; Langton, 2018b).

We explore whether and under what conditions ordinary people perceive a silent response when facing a hate speech incident as increasing the harm it creates, and how opposing speech may reduce that harm. Across two online experiments with UK participants using custom visual vignettes, we provide empirical evidence that bystanders' expression of opposition can modulate how harmful these incidents are perceived to be, but only as part of a collective response: one that is expressed by a substantial majority of bystanders, which suggests the existence of a social norm against hate speech.

Experiment 1 (N=329) shows that recognising the role played by silent or opposing bystanders who witness a hate speech incident depends on whether participants could take from the context the current social norm about how to respond to hate speech. In scenarios with three bystanders, participants recognised those who show opposition as reducing the harm created by the hate speech, while regarding those who remain silent as increasing that harm. However, in scenarios where the hate incident occurred in front of only one bystander, thus not allowing participants to recognise the social norm in place, they assessed hate speech incidents as equally harmful, regardless of whether the single bystander showed opposition or remained silent.

Experiment 2 (N=269) shows this is not simply a matter of numbers but rather one of norms: only unanimous opposition reduces the public perception of the damage created. Based on our results, we advance an empirical norm account: group responses to hate speech modulate its harm by indicating either a permissive or a disapproving social norm, which may guide bystanders' responses.

Our account and results show the need to complement individual responses with collective strategies (Chater & Loewenstein, 2022) against hate speech and other similar phenomena in which individual efforts seem not to suffice (e.g., climate change or global pandemics).

In Chapter 4, we return to our philosophical assumptions regarding hate speech to revise them in light of our empirical work. First, we interpret participants' stronger aversion to hate speech found in

Chapter 2 as the recognition that an intrinsic feature of hate speech acts is to perform various harmful actions through speech. These actions not only *cause* harmful effects, but also are harmful themselves and can target distinct people simultaneously. This is a result which other acts of hate cannot achieve and which is captured by folk intuitions. In addition, the need for collective responses and robust social norms in modulating hate speech harm, highlighted in Chapter 3, make us venture a new characterisation of hate speakers, challenging the idea that they are merely individuals or "lone wolves", and instead casting them as members of a group.

Accordingly, we defend the idea that it is a mistake to consider hate speech purely from the perspective of individual rights to free expression and discussion. The mistake is to treat harmful group speech with normative — and thus social — implications as an individual matter. Once hate speech actions, whether actual or perceived, become accepted by the majority, they risk mutating into policy options, susceptible of being supported by economic or populist lobbies interested in ascension to power through social confrontation. Although further developing this idea is largely beyond the scope of the present dissertation framework, we argue that harmful group speech with policy aspirations against disfavoured and minoritarian groups should not be granted freedom of speech protection under equal conditions alongside individual speech.

Furthermore, we highlight the pressing need to challenge the assumption that ordinary people are lenient toward hate speech and other forms of hatred and intolerance. It is crucial to recognise and harness people's capacity to discern the harm caused by such practices

and involve them in collective responses against hate speech. By doing so, we can effectively counter harmful discourses and foster tolerance and respect for diversity, thereby promoting peaceful coexistence within diverse societies.

Finally, Chapter 5 provides a comprehensive summary of our research, highlighting its key findings and their implications.

Zusammenfassung

Die vorliegende Dissertation ist in fünf Kapitel unterteilt. Einleitend werden in Kapitel 1 die Hassrede, der von ihr verursachte Schaden und ihre Zuhörer charakterisiert. Im Rahmen unserer Studie vertreten wir die Idee, dass Hassrede ein schädlicher Mechanismus ist, welcher in intergruppalen Auseinandersetzungen um soziale Dominanz verwendet wird (Charles-Toussaint & Crowson, 2010; Duckitt & Sibley, 2017; Hoover et al., 2021). Zudem zielt sie auf Personen basierend auf ihrer eigentlichen oder wahrgenommenen „Rasse“, Farbe, Abstammung, Nationalität, ethnischer Herkunft, Alter, Behinderung, Sprache, Religion, Geschlecht, Gender, sexueller Orientierung und anderer Identitätsmerkmale (Mihajlova et al., 2013). Deshalb bedroht ihre Verbreitung die Koexistenz innerhalb von vielfältigen Gesellschaften, wo Menschen mit verschiedenen Identitäten und Hintergründen zusammenleben.

In Anlehnung an die Sprechakttheorie, welche „illokutionäre“ Sprechakte als solche definiert, welche etwas schaffen indem man etwas sagt (z.B. jemanden heiraten indem man „Ja, ich will“ sagt), stellen wir Hassrede als solchen Akt dar (Langton, 2018a; Maitra, 2012), und vertreten die Idee, dass ihre Macht, anderen zu schaden, direkt mit ihrer Fähigkeit zusammenhängt, verschiedene schädliche Handlungen durch Sprache zu begehen.

Mit den Worten „Wir wollen euresgleichen hier nicht!“, „Geh zurück nach Hause!“, „Hör auf unser Land zu Islamisieren“, neben anderer paradigmatischer Hassrede, führen wir mehrere Handlungen aus: Wir stufen Minderheiten oder benachteiligte Gruppen als

minderwertig ein (Langton, 2012; Langton, 2018a; Langton, 2018b; Maitra, 2012), fordern direkte Bystander dazu auf, sich auf die Seite der Hassredner zu stellen, fordern Menschen zum Gehen auf, wenn sie nicht des Sprechers Standards von einem guten Bürger genügen (Fraser, 2023; Lepoutre 2021; Waldron, 2012), ermutigen Gleichgesinnte dazu gegen Zielgruppen vorzugehen, und führen manchmal mehrere dieser oder alle diese Handlungen gleichzeitig aus (Lewiński, 2021).

Besonders verteidigen wir die Idee, dass die meisten Hassreden nicht nur Schaden verursachen, sondern einen Schaden für die Zielgruppen darstellen (Langton, 2018a): Ob die Einstufung einer Gruppe als minderwertig, die Schuldzuweisung an Minderheiten für Umstände, die außerhalb ihrer Kontrolle liegen, oder die Aufforderung an benachteiligte Gruppen zu gehen oder eine besondere Behandlung dieser durch Hassrede, all dies verletzt Menschen (Langton, 2018a). Zusätzlich können diese Handlungen Gefühle von Demütigung, Hilflosigkeit, Isolation, geringem Selbstwertgefühl und Wut auslösen (Fattoracci & King, 2023; King et al., 2011), und Menschen psychologisch und physiologisch in verschiedenem Maße verletzen (Eisenberger, 2015).

Zudem charakterisieren wir Hassrede als stark kontextabhängig (Moreno, & Pérez-Navarro, 2021) Diese Eigenschaft trägt dazu bei, dass sie als an verschiedene Personen gerichtet wahrgenommen wird, wodurch ihre Zuhörerschaft und folglich die Personen, denen sie schadet, von den direkten Zielpersonen auf Bystander und die Gesellschaft als Ganzes ausgeweitet werden. Abhängig vom Kontext kann ein Zuhörer von Hassreden jeder von uns sein. Ungeachtet dessen, als Zuhörer spielen wir alle eine relevante Rolle bei der Bestimmung, Modulation und Bekämpfung der Handlungen, die ein Sprechakt ausüben kann. Deshalb sollte ein guter Ausgangspunkt, um dieses Phänomen anzusprechen, uns einbeziehen, um als potentielle Zuhörer zu identifizieren ob und in welchem Ausmaß wir eine Hassrede als schädlich wahrnehmen und unter welchen Bedingungen eine Reaktion dazu beiträgt ihren Schaden zu verringern.

Im Rahmen eines experimentellen Ansatzes zur moralischen Philosophie haben wir zur Bestätigung unserer Charakterisierung von Hassrede und ihrem Schaden zwei Forschungsarbeiten durchgeführt, um die Intuitionen gewöhnlicher Menschen zu diesen Themen zu überprüfen, welche in den Kapiteln 2 und 3 wiedergegeben werden.

Kapitel 2 reproduziert einen im Mai 2023 in Scientific Reports (Springer Nature) veröffentlichten Registered Report welcher gemeinsam mit Prof. Dr. Ophelia Deroy verfasst wurde. In dieser Studie testeten wir, ob Menschen eine intrinsische Abneigung gegen verbale Gewalt im Vergleich zu anderen Arten von Hasshandlungen (nonverbale, körperliche Handlungen) haben, und in welchem Ausmaß. Da Bystander nur selten Vorfälle von Hassreden melden, und in Anbetracht des rechtlichen, theoretischen und sozialen Zögerns, sie zu bestrafen, haben wir die Hypothese aufgestellt, dass Menschen bei Hassreden nachsichtiger sind als bei nonverbalen Hasshandlungen, wenn die Absichten und Folgen die gleichen sind. Wir führten ein Experiment mit 1309 britischen Bürgern durch, die Beschreibungen von verbalen und nonverbalen Vorfällen lasen, die auf identische hasserfüllte Absichten zurückgingen und welche die gleichen Folgen hatten. Wir fragten sie, wie viel Strafe der Sprecher (Täter) erhalten sollte, wie wahrscheinlich es ist, dass sie einen solchen Vorfall anzeigen würden, und wie schädlich diese Handlungen waren.

Die Ergebnisse widersprachen unseren zuvor registrierten Hypothesen und den Vorhersagen der dualen Moraltheorien, welche besagen, dass Absicht und schädliche Folgen die einzigen psychologischen Determinanten für Bestrafung sind. Stattdessen bewerteten die Teilnehmer verbale Hassvorfälle durchwegs als strafwürdiger und anzeigenswerter, und als schädlicher als nonverbale Vorfälle. Dieser Unterschied blieb auch dann bestehen, wenn wir die Szenarien so veränderten, dass die Teilnehmer wussten, dass die Zielpersonen keine negativen Konsequenzen erlitten (z. B., wenn die Zielperson taub war und deshalb die rassistische Bemerkung nicht hören konnte).

Wir erklären diesen Unterschied mit dem Konzept der Handlungs-Aversion (Miller et al., 2014), im Gegensatz zur Outcome-Aversion, suggerierend, dass Laienbeobachter etwas wahrnehmen, das der Sprache inhärent ist und sie dazu veranlasst, diese als schädlicher und bestrafungs- und anzeigenswerter zu bewerten im Vergleich zu anderen nonverbalen Hasshandlungen, unabhängig von ihren Konsequenzen. Später, in Kapitel 4, interpretieren wir das intrinsische Merkmal, das von den Intuitionen der Menschen in Bezug auf die Hassrede erfasst wird, als die Fähigkeit der Sprache, zu Schaden indem man etwas sagt (d.h., einen Schaden darzustellen und ihn nicht nur zu verursachen), und verschiedene Personen gleichzeitig anzugreifen (z. B. direkte Zielpersonen und zufällige Bystander), die auch dann bestehen bleibt, wenn es den Sprechern nicht gelingt ihre direkten Ziele zu schädigen.

Kapitel 3 reproduziert einen Artikel, der bei der wissenschaftlichen Zeitschrift Humanities and Social Sciences Communications eingereicht wurde und derzeit begutachtet wird. Der gemeinsam mit Prof. Dr. Ophelia Deroy, Dr. Justin Sulik und Mr. Clemens von Wulffen verfasste Artikel untersucht die Wahrnehmung der Rolle von stillen oder opponierenden Bystandern bei der Verringerung des durch Hassreden verursachten Schadens. Wir gehen von zwei weit verbreiteten Annahmen aus: einer vorherrschenden passiven Haltung gegenüber Hassrede und der Berücksichtigung von Gegenreaktionen von Bystandern in der Konfrontation mit einem Fall von Hassrede als hilfreich in der Schadensminimierung, etwas, das von Regierungsbehörden, Soziologen und Philosophen betont wurde (Ayala & Vasilyeva, 2016; Langton, 2018b).

Wir untersuchen, ob und unter welchen Bedingungen gewöhnliche Bürger eine stille Reaktion auf einen Vorfall mit Hassrede als Verstärkung des Schadens wahrnehmen, und wie eine Gegenrede ihn verringern kann. In zwei Online-Experimenten mit Teilnehmern aus Großbritannien, bei denen maßgeschneiderte visuelle Vignetten verwendet wurden, konnten wir empirisch nachweisen, dass die Äußerung des Widerstands von Bystandern die Wahrnehmung der Schädlichkeit solcher Vorfälle beeinflussen kann, allerdings nur als Teil einer kollektiven Reaktion: einer Reaktion, die von einer erheblichen Mehrheit der Bystander geäußert wird, was auf das Vorhandensein einer sozialen Norm gegen Hassreden schließen lässt.

Experiment 1 (N=329) zeigt, dass das Erkennen der Rolle von schweigenden oder opponierenden Bystandern, die einen Vorfall mit Hassreden beobachten, davon abhängt, ob die Teilnehmer dem Kontext die aktuelle soziale Norm darüber entnehmen konnten, wie auf Hassreden zu reagieren ist. In Szenarien mit drei Bystandern erkannten die Teilnehmer, dass diejenigen, die eine Gegenreaktion zeigen, dazu beitragen, den durch die Hassrede entstandenen Schaden zu verringern, während diejenigen, die schweigen, den Schaden vergrößern. In Szenarien, in denen sich der Hassvorfall vor nur einem Bystander ereignete, was den Teilnehmern folglich nicht erlaubte, die geltende soziale Norm zu erkennen, bewerteten sie Vorfälle mit Hassreden als gleichermaßen schädlich, unabhängig davon, ob der einzige Bystander eine Gegenreaktion zeigte oder schwieg.

Experiment 2 (N=269) zeigt, dass dies nicht einfach nur eine Frage der Anzahl ist, sondern eher eine der Normen: nur einstimmiger Widerspruch reduziert die öffentliche Wahrnehmung des erzeugten Schadens. Auf der Grundlage unserer Ergebnisse schlagen wir eine empirische Regel vor: Gruppenreaktionen auf Hassreden modulieren den Schaden, indem sie entweder eine nachsichtige oder eine missbilligende soziale Norm anzeigen, die die Reaktionen der Bystander leiten kann.

Unsere Darstellung und unsere Ergebnisse zeigen die Notwendigkeit, individuelle Reaktionen mit kollektiven Strategien (Chater & Loewenstein, 2022) auf Hassreden und ähnliche Phänomene zu ergänzen, bei denen individuelle Bemühungen nicht auszureichen scheinen (z. B. Klimawandel oder globale Pandemien).

In Kapitel 4 kehren wir zu unseren philosophischen Annahmen über Hassrede zurück, um sie im Lichte unserer empirischen Resultate zu überarbeiten. Erstens interpretieren wir die stärkere Aversion der Teilnehmer gegenüber Hassreden wie in Kapitel 2 als die Wahrnehmung, dass es ein intrinsisches Merkmal von Hassreden ist, verschiedene schädliche Handlungen durch Sprache auszuführen. Diese Handlungen verursachen nicht nur schädliche Effekte, sondern stellen selbst einen

Schaden dar und können verschiedene Menschen gleichzeitig schädigen. Das ist ein Ergebnis, welches andere Hasshandlungen nicht erreichen können, und das vom Volksempfinden erfasst wird. Darüber hinaus lässt uns die Notwendigkeit für eine kollektive Reaktion und für robuste soziale Normen zur Modulation des Schadens von Hassrede, hervorgehoben in Kapitel 3, eine neue Charakterisierung von Hassrednern wagen, die Idee hinterfragend, dass diese lediglich Individuen oder „einsame Wölfe" sind, und betrachten diese stattdessen als Mitglieder einer Gruppe.

Dementsprechend verteidigen wir die Idee, dass es ein Fehler ist, Hassrede ausschließlich aus der Perspektive von individuellen Rechten auf freie Meinungsäußerung und Diskussion zu betrachten. Der Fehler ist es, schädliche Gruppensprache mit normativen — und damit sozialen — Implikationen als individuelle Angelegenheit zu behandeln. Sobald die durch Hassrede durchgeführten Handlungen, ob tatsächlich oder empfunden, mehrheitlich akzeptiert werden, besteht die Gefahr, dass sie zu politischen Optionen mutieren, die von wirtschaftlichen oder populistischen Lobbys unterstützt werden können, welche daran interessiert sind, durch soziale Konfrontation an die Macht gelangen. Auch wenn die weitere Entwicklung dieser Idee den Rahmen dieser Dissertation sprengen würde, argumentieren wir, dass schädliche Gruppenäußerungen mit normativen Ansprüchen gegenüber benachteiligter und minoritärer Gruppen nicht unter gleichen Bedingungen wie individuelle Äußerungen durch die Redefreiheit geschützt werden sollten.

Darüber hinaus betonen wir die Dringlichkeit, die Annahme von Nachsicht von gewöhnlichen Bürgern gegenüber Hassrede und anderen

Formen von Hass und Intoleranz in Frage zu stellen. Stattdessen müssen wir ihre Fähigkeit nutzen, den von Hassrede verursachten Schaden zu erkennen, um ihn kollektiv zu bekämpfen. Das bedeutet, wir sollten Initiativen fördern, welche schädlichen Diskursen entgegenwirken und Werte von Toleranz und Respekt für Diversität stärken, zu Gunsten einer friedlichen Koexistenz innerhalb verschiedener Gesellschaften.

Abschließend fasst Kapitel 5 die wichtigsten Ergebnisse unserer Untersuchung zusammen.

Chapter 1

Characterising hate speech and its audience.

1.1. Doing harm with words.

"Sticks and stones may break my bones, but words will never hurt me," states an old saying, much-quoted across the centuries. But does it remain in force currently? Do we still accept as fact that speech cannot harm us?

By "harmful" speech, we will not allude to all cases of offensive or abusive language that, while being unfortunate, despicable, and morally reproachable, are expressions of dislike against a particular individual, even though they can distress their targets (Jay, 2009). We will restrict our attention to so-called hate speech. Despite ideas about the freedom of expression principle varying widely across regions and nations (Wike & Simmons, 2015), we argue that citizens of established democracies must learn to live with those language overruns whilst being aware that most hate speech does, in fact, harm people, reinforcing social confrontation that typically ends in violence against minorities and disfavoured groups, and fracturing our societies (Benesch, 2013; Leader Maynard & Benesch, 2016).

Hate speech does not have an agreed upon definition. Different kinds of speech can be considered hate speech and hurt people in various ways. In our present work, we are interested mainly in how hate speech erodes the coexistence within diverse societies, where people from various backgrounds and perspectives live together. Thus, we shall characterise hate speech as one that demeans a social group, excludes it, or directs said group to leave due to their actual or perceived "race", colour, descent, national or ethnic origin, age, disability, language, religion, sex, gender, sexual orientation and other similar features (Mihajlova et al., 2013).

We want to highlight that the features mentioned do not belong to a single individual but are typically shared by a group. Therefore, we defend the idea that hate speech targets groups, not individuals. When hate speech seems to target a single individual, it does so based on the (target's/victim's) actual or perceived membership of a minority or disfavoured group against which the hate speaker is biassed.

Additionally, we argue that hate speech not only carries further damaging consequences for its targets (e.g., feelings of humiliation or isolation). It is harmful itself. Being ranked as inferior, excluded, or directed to leave the community based on who we are, whether that exclusion is based on our self-identity or the assumptions of others, constitutes a real harm to ourselves as well as our communities. (Langton, 2012, 2018b; Gelber and McNamara, 2016; Maitra, 2012). Moreover, we characterise hate speech as an action, not just a linguistic expression.

1.2. Theoretical framework.

The present dissertation approaches hate speech from the perspective of speech act theory. Introduced by J. L. Austin, this theory describes three kinds of speech acts: locutionary, illocutionary and perlocutionary. Locutionary acts are those of saying something: they relate to the linguistic content and its reference (e.g. saying “the sky is blue”). Illocutionary acts involve doing something by saying something (e.g. marrying someone by saying “Yes, I do”): they relate to the action performed (which may not coincide with what the speaker intends to achieve, as will be explained later). Perlocutionary acts are those which cause something by saying something (e.g., amusing someone by telling a joke): they relate to the consequential effects that the speech has on the hearers, extended audience, or even further (Austin, 1962).

Thus, in saying, "I do not want soup for lunch" or "You did a great job!" we *refuse* to take soup, or *congratulate* a friend. If, when seeing an Asian-looking person, we yell at her, "We do not want your kind here!" or "Chinese are making our country sick!" regardless of whether we do it driven by hatred and intolerance, by fear of catching the COVID-19 virus or by any other intention, we are also *directing* Asian-looking people to leave the country or *blaming* them. Therefore, we argue that hate speech can be approached as a particular kind of speech act: an illocutionary speech act (Langton, 2012; 2018a; Lepoutre, 2017, 2023; Maitra, 2012).

Speech act theory describes an illocutionary speech act as a tool for interpersonal interaction used by a speaker to perform a particular action (Austin, 1962; Bach & Harnish, 1979). It is in this sense that hate speech not only directs others to leave the country, to set aside their religion or to change their gender orientation, but it also "subordinates an entire social group, places people in hierarchies, deprives them of powers and rights, and legitimates certain treatment of them" (Langton, 2018a).

Moreover, hate speech can harms us in an "assaultive" way when a group orders us to leave the country on a bus, using homophobic epithets, or in a "propagandistic" way when we hear a politician degrading asylum seekers in the news (Langton 2012 and Gelber & McNamara, 2016 as cited in Anderson & Barnes, 2022). In those cases, hate speech can order a disfavoured group to leave the country whilst dictating to bystanders how things are and what they are permitted to think, love or hate. Depending on the context and the audience present, hate speech can do both things simultaneously (Lewiński, 2021).

Hate speech constitutes harm and "enable[s] the enactment of norms and hierarchies that are socially real" (Langton, 2018a, p. 136; Langton, 2018b). In addition, it can cause harmful and long-lasting effects like feelings of humiliation, helplessness, isolation, low self-esteem or anger (Fattoracci & King, 2023; King et al., 2011). However, as it will be explained in Chapter 2 when due to contextual features or even by luck, hate speakers do not manage their speech to cause hurting effects on direct targets, it remains perceived as harmful by third-party observers.

The harm of hate speech occurs in different ways, psychologically and physiologically, and in various degrees. Empirical evidence "supports the hypothesis that physical and social pain relies on shared neural and neurochemical substrates" (Eisenberger, 2015, p. 623), suggesting that social rejection and exclusion hurt more than in a metaphorical way (See Eisenberger, 2015 for a review of those findings). Studies conducted with direct targets and bystanders also report the harm verbal attacks create (Nielsen, 2009). However, harm through speech has been cast into serious doubt by the assumption that what is harmful to one individual may not necessarily be seen as such by others, and by the further assumption that people do not consider hate speech as harmful as other exhibitions of hate or rejection. As the context in which incidents of hate speech happen plays an essential role on to what extent those incidents harm direct targets and bystanders, it puts serious limits on the generality of the studies conducted to clarify these aspects.

Furthermore, hate speech acts are usually performed in daily-life interactions between strangers or in public communication, where speakers often have to manage without knowing their audience or their expectations and circumstances (Camp, 2018; Moreno & Pérez-Navarro, 2021). There, our cultural, social, political, economic and personal circumstances, like nationality, profession or tone of voice, play a crucial role in determining what we do with our words (Austin, 1962; Green, 2021; Searle, 1979). If we are white-skinned and greet an African-American colleague using the "N-word", our audience might have reasons to attribute to us a negative attitude towards African-Americans (Moreno & Pérez-Navarro, 2021). And this can happen without us being aware of it.

Let us elaborate on this with an example: we enthusiastically ask our 10-year-old son to choose a colour for his bedroom walls, he chooses a pink colour, and we feel uncomfortable and hesitant before accepting it, we can be aware that the discomfort comes from a sexist bias. A sexist bias can be accessible to us or operate outside our awareness but it still influences our actions (Basford et al., 2014).

Moreover, we can despise someone intentionally or unintentionally: calling an African American work colleague's hair "exotic" intending to compliment them allows others to attribute to us a racist bias. Ignoring a client in a store because of her accent can be intended to cause harm. In both cases, our actions are racist, regardless of our actual intentions.

In addition to such context-dependence, the intrinsic nature of speech, as essentially social and interactive, does not give the speakers absolute control over which actions they perform in their speech. The speech act theory recognises that the speech acts we perform are the result of an interactive process between the speaker and the audience. When the speaker performs certain actions through her words, the audience also has a role in "fixing the import, success, influence, and social life" of the resulting speech act. (Kukla, 2023).

If someone attacks us with sticks and stones, we can always take anything that comes to hand and defend ourselves by using it as a shield. Then the question arises: if saying is doing, is there a possibility to defend ourselves from hate speech?

Our answer is yes. We can do so in different ways: from engaging in counterspeech and questioning the harmful presuppositions hate speech introduces in the social dialogue (Ayala & Vasilyeva, 2016; Cepollaro et al., 2023; Langton, 2018b) to denouncing a hate speech incident to competent authorities (in countries with laws against hate speech), or by supporting its targets (e.g., giving them platforms to raise their voice against hate speech) to participating in electoral processes (e.g., voting against political options that spread rejection, hatred and discrimination against minorities or disfavoured collectives). In the success of those strategies, bystanders who witness a hate incident and the entire society are crucial.

This understanding of hate speech as an (illocutionary) action or actions allows us not to restrict specific locutions (linguistic content), which semantic value can mutate, nor in the (perlocutionary) effects hate speech can cause, which can be avoided depending on the context (e.g., could not be heard) but to the harmful actions performed by words as inherently harmful. Moreover, considering hate speech acts as "embedded in social communities, relationships, and ecosystems" (Kukla, 2023) is essential because the targets and any of us, as an audience, can respond against it, not silencing the words themselves but impeding the actions they perform.

Therefore, identifying who makes up the audience of hate speech becomes crucial.

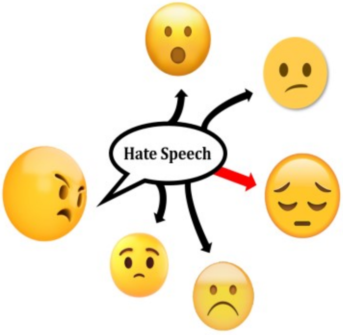
1.3. The audience of hate speech.

As we already mentioned, the social and interactive nature of a speech act places beyond the speaker's control how the audience perceives what is said and which actions are performed through her words. Moreover, it puts out of her control two further aspects: (a) Who makes up the audience and (b) whether this audience perceives the speech as coming from a group member or an individual. We will develop the first idea hereafter, leaving the second for Chapter 4, where we characterise the hate speaker.

Who, then, makes up the audience of hate speech? As hate speech points to disfavoured collectives or minorities, demeaning them based on their identity marks, it has been shown that when a hate speech act targets a member of a disfavoured group or minority, the whole group receives the action (Gelber & McNamara, 2016). Research done with hate incidents (including hate speech) has shown that hearing someone who shares with us identity features suffered a hate attack might cause us to experience high levels of distress and feelings of humiliation, helplessness, isolation, low self-esteem, anger, fear and anxiety, to name but a few (Cook & Sheppard, 2018; Schmader et al., 2012; Swim et al., 2001, 2003). Therefore, when a hate speech act is performed, its audience lies beyond the immediate individual addressed: it is a whole group. This group-targeting occurs irrespective of hate speakers' intention and regardless of their awareness, as we mentioned above.

Imagine that during the COVID-19 pandemic you took a bus and saw an Asian-looking person. You then yelled at her, "Get out! Don't bring the virus here!" Even if you argue that you just wanted to make that concrete individual leave the bus because you were scared about getting sick and that you do not intend to derogate all Asian-looking people, certainly your speech might be taken as targeting all Asian-looking people, ordering them to leave and blaming them for spreading the coronavirus, regardless of whether or not they are carriers of the virus.

Figure 1. The audience of hate speech.



Furthermore, as daily interactions between strangers are often set in public spaces, by performing a hate speech act, we might perform multiple actions directed to different members of our audience: If in a crowded city bus, when looking at a passenger wearing an Abaya, we yell at her, "Stop Islamization of our country!" we might be performing several speech acts: ordering Muslims to leave, directing bystanders how not to dress or encouraging like-minded fellows present to take action against Muslims. Lewiński (2021) calls these "plural illocutionary speech acts".

In this way, bystanders who witness a hate speech incident might also become its audience regardless of whether they are members of the target group. Moreover, they might also be harmed by hate speech in different ways. Empirical research has confirmed that exposure to such speech harms bystanders: witnessing repetitive verbal mistreatment and abusive discrimination affects both bystanders and direct targets in similar physiological and psychological ways (Janson & Hazler, 2004; Janson et al., 2009). The exposure to hate speech has been linked with more significant desensitisation to demeaning expressions (Greenberg & Pyszczynski, 1985) and a decreasing sympathy for the targets of hate speech (Leets 2001, Greenberg & Pyszczynski 1985 and Carnagey et al. 2007 as cited in Soral et al., 2018), which reinforces outgroup prejudice, eroding social coexistence.

Thus, any of us can become part of the audience of a hate speech act; as such, the harm it creates can reach us, from targets and bystanders to whole communities (Anderson & Barnes, 2022).

1.4. Methodological approach to hate speech harm.

Currently, the role of the audience in an illocutionary act, such as hate speech, is the focus of debate (McDonald, 2022; Schmitz &

Townsend, 2020). Some scholars argue that the potential actions a speaker performs through words is determined by the speaker alone, by her expression of a particular communicative intention (Alston 2000, Bird 2002, Harris 2019a; Jacobson 1995 as cited in McDonald, 2022); others maintain that the hearer's uptake—the hearer's "understanding of the meaning and the force of the locution" (Austin, 1962)—is essential in determining the illocutionary force of a speech act. Depending on the context, this could even prevail when the hearer's interpretation contradicts the speaker's intention (Austin 1962; Hornsby and Langton 1998; Langton 1993; McDonald 2021 as cited in McDonald, 2022).

Finally, some hold that the illocutionary force is the product of a collaboration between the hearer and the speaker. Once the speaker utters something, the hearer communicates to the speaker that she interprets the utterance as having a certain force, and likewise, the speaker communicates to the hearer that she accepts the hearer's interpretation" (McDonald, 2022). This collaboration occurs through a signalling process facilitated by conventional responses which we expect the hearer to provide, like smiling to someone who says something about our hair to let her know that we receive those words as a compliment (McDonald, 2022; Sbisà, 1984; 2001).

But hate speech, as we have characterised it, performs actions that harm people. Thus, targets would unlikely collaborate with speakers in confirming the performance of a hate speech act. Similarly, hate speakers would unlikely let their targets have a prevailing role in determining the illocutionary force of their speech.

In addition, as we will address in Chapter 2, acknowledging that hate speech harms people has led many countries to legislate against it, putting the most extreme cases in a court of law. In those cases (which remain exceptional), the speakers will undoubtedly deny having performed a hate speech act. Of course, there will be cases where an extreme-right supporter proudly confirms the hate intent, but that will not be the case in the vast majority of hate incidents, where hate speakers are at least partially unaware of the harmful potentiality of their words or will deny any hateful, demeaning or discriminatory intent to avoid sanctions. Moreover, as hate speech harms people by degrading or excluding them, refusing the normative power of a speech act by an uncooperative uptake might be reasonable and even ethical (Kukla, 2023).

So, how should we study the harm hate speech creates? From the third-party perspective we will explore in which circumstances and contexts ordinary people perceive hate speech as harmful; are keener to counter it; and perceive showing opposition as efficient. Studying how hate speech functions in the eyes of ordinary citizens might better inform public policies directed to them that aim to change the apparent leniency towards hate speech (Barhight et al., 2017; Cook & Sheppard, 2018; Urschler et al., 2015; Wenik, 1985), helping to identify the most effective strategies to counter hate speech (Gulker et al., 2013).

As we mentioned, the audience of hate speech comprises members of a target group but also out-group people exposed to hate speech when hearing the news, walking the streets, or waiting for a bus, who can nonetheless respond to hate speech. Any of us ordinary people can be the audience of hate speech. Thus, we all can refuse to cooperate with the performance of a hate speech act. Are we keen to do so? Do we consider our responses to help to reduce hate speech harm?

It has been shown that ordinary people perceive some hate speech as highly severe and harmful to the group attacked (Leonhard et al., 2018). However, the perception of severity alone does not increase people's intention to counter hate speech. It requires that they feel personally responsible for intervening (Leonhard et al., 2018). Under which conditions does this occur?

In the last decade of the twentieth century, a new methodology to test philosophical theories emerged as a branch of the larger experimental philosophy. So-called "experimental moral philosophy" mainly studies ordinary people's moral intuitions, judgments and behaviours using experimental methods from the cognitive sciences. The studies involve data collection that is analysed afterwards using statistical tools to "substantiate, undermine, or revise philosophical theories" (Alfano et al., 2022).

In the present dissertation, we follow that methodology and contrast our philosophical assumptions about hate speech and the harm it creates within the intuitions of ordinary people. We will explore how people perceive the harm created by hate speech, in contrast to how they perceive the harm created by similar kinds of mistreatment (Skarlicki & Kulik, 2004; Skarlicki et al., 2015), aiming to inform public policies against hate speech in real-life situations and in online settings, where hate speech is a massive problem (Agatston et al., 2007; Kim, 2021; Lytle et al., 2021; Mishna et al., 2010; Rovira et al., 2021).

Following this methodology, after characterising hate speech, its harm and its audience in Chapter 1, we will report three experimental studies in Chapters 2 and 3, in which we explore people's moral intuitions about hate speech, the harm it creates and the role played by the audience in opposing it. Finally, in light of our findings, we will revise our initial assumptions and propose a characterisation of the hate speaker in Chapter 4.

Chapter 2

Perceiving hate speech's harm: The illusion of ordinary people's leniency against hate speech[1].

2.1. Introduction

Should we prosecute people for their words and not just their deeds? A robust liberal tradition inspired by John Stuart Mill's theories argues that Freedom of Speech, as a constitutional right, is granted with superior protection against State regulatory interference, despite possible conflicts with morality (Mill, 1859). However, most European countries have introduced hate speech regulations to reframe that privilege, sanctioning verbal and nonverbal hate attacks similarly when they share analogous severity and degrading intention and create comparable consequences (Barendt, 2009; Belavusau, 2012; Boyne, 2010). Legal theorists are not the only ones to disagree on the extent to which speech should be punished: philosophers, legislators, politicians, activists, and citizens are highly divided on the issue. The persisting reluctance to sanction hate speech, at least as much as other hate crimes, raises a question for moral psychologists:

Is this lenience ingrained in our moral dispositions, and if so, how?

[1] **Important note**: The content of the present chapter reproduces in its entirety a paper co-authored with Prof. Dr. Ophelia Deroy, published as a Registered Report:
Zapata, J., Deroy, O. Ordinary citizens are more severe towards verbal than nonverbal hate-motivated incidents with identical consequences. *Sci Rep* 13, 7126 (2023). https://doi.org/10.1038/s41598-023-33892-8

CHAPTER 2. PERCEIVING HATE SPEECH`S HARM: THE ILLUSION OF ORDINARY PEOPLE'S LENIENCY AGAINST HATE SPEECH

Linguists and comparative psychologists argue about fine distinctions between verbal and nonverbal actions. Still, most accept that verbal expression is unique amongst other possibly communicative behaviours, such as gestures, facial expressions, or bodily actions: It is the most effective and nuanced way to express mental states, including feelings, abstract ideas, or hypotheticals, and it may be uniquely able to communicate complex information. From a legal perspective but also based on speech uniqueness, the American Constitution's First Amendment (1791), the Declaration of the Rights of Man and the Citizen (1789) and the Universal Declaration of Human Rights (1948) enshrined the Freedom of Speech principle to protect the intrinsic value of speech as a means for free expression and discussion [*Due to the commonness of the phrase "Freedom of Speech", we will refer to "speech" in what follows, but mean this in a way that includes all language, whether spoken, signed or printed. We also intend "nonverbal" to exclude all uses of such linguistic forms.*].

However, beyond those historical and traditional considerations, this distinction between verbal and nonverbal actions is not a foregone conclusion. Speech can have consequences and cause harm no less than nonverbal actions do (Alexander, 1983; Bayles, 1986; Dworkin, 1977; Redish, 1984; Scanlon, 1972; Schauer, 1982, 1993, 2015) - granting that harm definition does not reduce to the classic notion of pain as tissue damage (Cohen, 2018; Raja et al., 2020). The comparison between the negative impact caused by verbal and nonverbal actions does not depend here on how much neural overlap there is between the painful experience caused by social rejection or exclusion and physical damage to the body (Eisenberger, 2015). Negative consequences for the individual are usually captured by a folk concept of harm, deployed in moral judgment, which is certainly broader than the concept of "pain" in a sense that is tied to a specific neural activation (Schein & Grey, 2018). In addition, the legal concept of harm, distinguished from the mere offence, is broader than actual or possible tissue damage. As argued amongst legal scholars, harm can also include negative consequences for one's well-being caused by speech (Petersen, 2016), expanding the notion of harm used by Mill nearly two centuries ago (Bell, 2021).

Depending on the context, harm caused by speech could be as permanent, long-lasting and intense as that caused by nonverbal actions (Delgado, 2013; Waldron, 2012). Studies conducted with victims of hate speech and verbal abuse have found high levels of distress and severe psychological damage due to exposure to humiliating, demeaning or discriminatory speech (Nielsen, 2009; Wabnitz et al., 2012; Walters, 2014a; Weinberg & Nielsen, 2017). Having made this clear, we should ask whether third-party observers would evaluate verbal and nonverbal hate attacks similarly when both share the same intention to harm others and similar consequences, and if so, why.

The need for answers to these questions is clear when the evidence about moral attitudes towards harmful speech remains equivocal and non-specific. Relevant nationwide surveys on either side of the Atlantic continuously report many citizens who oppose hate speech bans despite believing that harmful discourses are morally unacceptable (Ekins, 2017; Kellner, 2012; Wike & Simmons, 2015). In the same line, college-student survey respondents broadly support free speech but increasingly favour restrictions on discourse that targets minority groups (Gallup, 2020; Naughton et al., 2017). However, looking at relevant official reports on hate speech, the high proportion of under-reported incidents is striking (Githens-Mazer & Lambert, 2010; Home

Office, 2019; OSCE/ODIHR, 2014; Spanish Ministry of Interior, 2019; UK Home Office, 2016; United States Department of Justice, 2018). Does this discrepancy between the increasing readiness to recognise speech as a crime and the lack of reporting hate speech incidents to authorities partly derive from a feature of our moral psychology? Are we less inclined to punish and denounce those who use words rather than physical actions to harm others, and if so, why?

Looking at the related literature in moral psychology, while moral evaluation and punishment rest on cognitive and emotional processes (Darley, 2009; Greene et al., 2001; 2004; Greene, 2007), dominant dual-system theories frame such evaluation as weighing the intention and the consequences of the action (Cushman, 2008; Gino et al., 2008a; Greene et al., 2009; Young et al., 2007). The exact equivalence in moral condemnation judgement is predicted by theories that argue that our punishment heuristics are driven mainly by outcomes (Prochownik & Cushman, 2019). Granting that verbal (hate speech) and nonverbal hate attacks share similar demeaning and harmful intentions, the explanation from the dominant dual-system theories needs to be that people evaluate the consequences of verbal and nonverbal attacks differently.

Against this consequentialist prediction, we reckoned that the comparative leniency toward hate speech comes not from minimising its harmful consequences but from the tendency to see verbal actions involving speech as inherently less morally negative than nonverbal actions involving the body. Our prediction here extends and complements a range of findings showing that moral evaluation and punishment is determined not only by intentions and consequences but also by associations with some properties of the actions themselves and people's aversion to them (Cushman et al., 2012; Cushman, 2013; Hannikainen et al., 2014; Miller & Cushman, 2013; Miller et al., 2014).

Miller, Cushman and Hannikainen define this "action aversion" as one's aversion to intrinsic action's properties. Together and separately, they found robust support for its importance in first-person and third party moral evaluations (Miller et al., 2014). They also showed that action aversion can predict harm condemnation in the context of moral dilemmas, where an affective response to victim suffering (outcome-aversion) cannot (Miller & Cushman, 2013). For instance, although the moral status of an action (e.g., lying is wrong) is usually assessed along with its expected consequences (e.g., lying will cause harm), some typically harmful behaviours (e.g., pushing a person off the footbridge in the so-called trolley problem) might be considered morally worse than atypically harmful ones (e.g., flipping a switch). Even when both lead to the same harmful consequences (e.g., the death of a victim). Experimental evidence also points in the same direction when it demonstrates that people are averse to performing pretend harmful behaviours (e.g., hitting a baby-doll or firing a toy gun towards a friend), even when they are aware of their harmlessness potentiality (Cushman et al., 2012; Cushman, 2013).

Following those findings, we conducted online a vignette-based study. On it, participants assessed verbal and nonverbal hate attacks in which derogatory intent and consequences for the victim (either negative or nonexistent ones) remained equal. We predicted that irrespective of perceiving that both attacks inflict similar harm on their victims, participants would punish and denounce more leniently those committed using words.

CHAPTER 2. PERCEIVING HATE SPEECH`S HARM: THE ILLUSION OF ORDINARY PEOPLE'S LENIENCY AGAINST HATE SPEECH

Given the lack of a consensus definition of hate incidents and hate speech in the literature, to delimit the scope of our study, we characterise hate actions as ones performed by a perpetrator with a degrading and discriminatory intention towards a victim. Based on a particular personal characteristic of the latter (race or ethnic origin, religion, gender, physical or mental conditions, among others). We deliberately avoid stories representing either extreme verbal violence (slurs and death threats) or nonverbal one (punching, beating or kicking), given that vignettes about such actions may distress participants overly, and they are rarely controversial regarding the obligation to denounce them. Moreover, hate speech incidents are less about insults and more about demeaning and discriminatory discourse targeting members of minority or disfavoured groups or identities. Such incidents convey a symbolic message to victims that they are unwelcome and unworthy of social respect (Perry & Alvi, 2012; Walters, 2014a). Therefore, in our study, participants assessed generic linguistic expressions that also convey harm (e.g. "Go back home!", or "We don't want your kind here!") and nonverbal degrading and discriminatory behaviours (e.g., spitting close to someone's feet, or stopping someone from sitting next to one).

We conducted the study with native English speakers from the UK. The British legal system has pioneered the implementation of strict hate speech regulations in Western Europe, dating back to the seventeenth century, and the UK is currently the European country that invests the most economical and human resources in combating hate speech and creating social awareness about verbal harm (Home Office, 2019; Rosenfeld, 2003). Therefore, if a more lenient approach to hate speech is confirmed even for people whose national legislation reinforces the idea that speech can be as harmful as nonverbal actions, this would bring more confidence for future replications.

Finally, the same concern of scope delimitation underlines the option in favour of testing only a single bias behind all hate-attack scenarios, religious hatred, which is remarkably consistent across countries (UK Home Office, 2016; United States Department of Justice, 2018). Since hate attacks are highly context-dependent, we seek this way to minimise the influence that a greater aversion to a specific bias may have on our results.

More importantly, in the UK, racially or religiously aggravated offences are, by definition, hate crimes, and just over half (53%) of hate crime' offences are recorded as one of these racially or religiously aggravated ones (Home Office, 2021). Therefore, race or ethnicity and religion have particular relevance to our study. In addition, they frequently overlap (Considine, 2017; Githens-Mazer & Lambert, 2010). However, while racial or ethnic bias could be linked with several victim profiles, just under half (45%) of religious hate crime offences were targeted against Muslims (Home Office, 2021). Thus, we have chosen to focus on religious hatred against Muslims rather than racial hatred for the present study, as the latter would potentially introduce greater unexplained variance. However, in further studies, we will seek to answer whether our results are replicable with different hate biases (e.g. hatred based on race) or groups of participants (e.g. Americans instead of British).

Our study provides two timely contributions to the literature: First, the extensive literature in moral psychology on blame and punishment mainly focuses on cases of physical pain (Cushman, 2008;

Gino et al., 2008a; Greene et al, 2009; Young et al., 2007); where damage is caused to someone's bodily integrity (e.g., killing, wounding) or monetary gains (Cushman et al. 2009). As words do not overtly or directly cause physical harm or affect economic gains, they are not considered by most literature on third-party punishment and moral judgments, with very few exceptions (e.g., Swim et al., 2003). And, to implement hate speech laws and assure their effectiveness, it is crucial to test whether we could or could not replicate the findings on blame and punishment of physical harm, on speech harm. Moreover, to our knowledge, our study is the first to experimentally apply the action-aversion principle to explain Hate Speech, filling this gap in the literature.

Second, several researchers have called for a better contextualisation of moral psychology (Schein, 2020). Our study contributes to this agenda by testing scenarios which avoid describing extremely violent aggressions, which are exceptional and focusing on more common demeaning and derogatory actions that victims often encounter. At the same time, it accounts for the role of identities in hate bias - moving away from the psychology of "raceless, genderless strangers" (Hester & Gray, 2020). While we are only testing one type of bias in this first study, as we focus on the psychological mechanisms underlying moral leniency towards hate speech attacks, follow-up studies should confirm and extend our findings to different hate biases, social identities, and groups of participants.

Against this background, we formulated two research questions:

The first one (**H1**): Do lay observers evaluate attacks that share the same hate intent and create similar negative consequences for the victim differently, depending on whether they are perpetrated through verbal or nonverbal actions? And going one step forward, the second (**H2**): Do lay observers evaluate attacks that share the same hate intent differently, depending on whether they are perpetrated through verbal or nonverbal actions, even when they create no consequences for the victim?

And, based on our pilot study, we hypothesized that -consequence and hate intent being the same- participants would be less inclined to punish and denounce hate attacks committed by verbal actions than nonverbal ones, irrespective of consequence type (negative or nonexistent). Additionally, we predicted that participants would rate that both attacks inflicted similar harm on their victims. Our pilot study (N = 171) provided confirmatory hypothesis results (See Stage 1 Registered Protocol): When both actions had the same hate-intent, participants assigned less punishment to verbal actions than to nonverbal ones, yet they rated both types of attacks as comparably harmful. This last finding was confirmed by Bayesian analysis (See Supplementary Information for pilot study details). Our pre-registered hypotheses are documented below:

H1. Participants will evaluate attacks that share the same hate intent and create similar negative consequences for the victim differently, depending on whether they are perpetrated through verbal or nonverbal actions: More leniently in terms of punishment and denunciation while considering both types similarly harmful.

H1. a: Participants will punish verbal hate actions less than nonverbal ones.

H1. b: Participants will be less likely to denounce verbal hate actions than nonverbal ones.

H1.c: Participants will rate verbal and nonverbal hate actions similarly harmful to the victim.

H2. Participants will evaluate attacks that share the same hate intent and create no consequences for the victim differently, depending on whether they are perpetrated through verbal or nonverbal actions: More leniently in terms of punishment and denunciation while considering both types similarly harmful.

H2. a: Participants will punish verbal hate actions less than nonverbal ones.

H2. b: Participants will be less likely to denounce verbal hate actions than nonverbal ones.

H2.c: Participants will rate verbal and nonverbal hate actions similarly harmful to the victim.

We defended that third-party lower moral condemnation of hate speech would be better explained by action aversion (response to intrinsic action's properties and their typically associated consequences, irrespective of their actual outcomes) than outcome aversion (response to action's consequences for the victim). Please note that we did not deny the role of outcome aversion in moral condemnation by no means. Instead, we defended the idea that when ordinary citizens face verbal and nonverbal hate attacks, the action aversion against verbal ones would be significantly lower independently of their consequences. We suggested that something intrinsic in words and speech, traditionally and historically linked with legitimate informational and cooperative purposes, would make people more likely to grant them special protected status and be more lenient towards their harm.

2.2. Experimental work: Methods.

2.2.1. Ethics Information

The study complied with all ethical regulations. Ethics approval was obtained from the local Ethics Committee at the LMU (ID-Number 131874 from 10.02.2022). Participants provided informed consent at the outset of the experiment, and all received relevant information about the research aim, procedure, duration, and compensation. They also were informed that no expected risk would be involved by taking part in the experiment and about the option of withdrawing from the study at any time without ensuing consequences. Participants were compensated with 1.20 pounds for 10 minutes of participation.

2.2.2. Study Description

The study is an online experiment based on the contrastive vignette method. A 2 x 2 mixed design was implemented with two independent variables (IV), the first, action type, as a within-subjects factor with two levels: verbal and nonverbal hate-attacks, and the second, consequence type, as a between-subjects factor with two levels: negative consequence and nonexistent consequence for the victim. In addition, participants' ratings of three dependent variables (DV) were collected in random order: Appropriate punishment, the likelihood of denouncing perpetrators to competent authorities, and the level of harm inflicted on the victim.

Participants, as lay-observers, contrasted situations where a character with the same degrading intention against a targeted victim performs either a verbal attack (using words) or a nonverbal (bodily) one. The description of the consequence for the victim in both cases was explicitly the same (See Supplementary Information section for Testing Materials).

Additionally, to better test the action aversion theory (which predicts that moral judgments are driven by one's aversion to intrinsic action's properties and their typically associated consequences, irrespective of their actual outcomes), we allocated participants randomly into two groups. Group A participants tested two experimental trials, verbal and nonverbal hate actions with identical negative consequences for the victim (e.g., as a consequence of the hate attack, the victim who suffered it stops using the bus line in which he was attacked). Group B participants also assessed two experimental trials, but this time, with no consequences for the targeted victims (e.g., we let participants know that the victim was deaf and could not hear the hate speech remark).

Participants in Group A were presented with six vignettes in random order: The two experimental trials, verbal and nonverbal, three distractors, and an attention check: 1) Verbal hate action based on religious hatred with negative consequences for the victim, 2) Nonverbal hate action based on religious hatred with negative consequences for the victim, 3) Distractor 1: Neutral action in a religious hatred context, 4) Distractor 2: Verbal hate action against meat eaters, 5) Distractor 3: Nonverbal hate action against environmental polluters, 6) Attention check. Participants in Group B were also presented with six vignettes in random order. Still, in this case, both verbal and nonverbal scenarios had no actual consequences for the victim: 1) Verbal hate action based on religious hatred with nonexistent consequences for the victim, 2) Nonverbal hate action based on religious hatred with nonexistent consequences for the victim, 3) Distractor A: Neutral action in a religious hatred context, 4) Distractor B: Verbal hate action against meat eaters, 5) Distractor C: Nonverbal hate action against environment polluters and 6) Attention check.

2.2.3. Testing Materials.

2.2.3.1. Experimental Vignettes.

All experimental vignettes shared a similar structure: Scenario setting (1, 2 or 3), description of the perpetrator's hostility towards a specific group (Muslims) to which the victim belongs (4), an opportunity to convert that hostility into action (5), the performance of a hate-attack (6 or 7), and a final outcome (8 or 9).

Participants were randomly assigned to one of the three scenario settings (1: Bus, 2: Train or 3: Supermarket) and, within that scenario, allocated into two groups, where the IV consequence type was manipulated: Participants in group A received two experimental trials (verbal and nonverbal) with negative consequences for the victim as the same final outcome (8). Participants in group B received the two experimental trials (verbal and nonverbal) with nonexistent consequences for the victim as the same final outcome (9). Aspects (4) and (5) remain broadly similar across vignettes. Finally, we manipulated the IV action type in the action's performance, being either verbal (6) or nonverbal (7).

An example of a nonverbal action was 'John looked Bilal in the eyes and spat on the ground next to him', while an example of a verbal action was 'John yelled at him, "Go home! Stop Islamization of our country!"'. (See Supplementary Information for the Study Testing Materials).

2.2.3.2. Distractors.

Three distractors were presented to participants. The inclusion of distractors in this study served two objectives: 1) To make the study goals less evident, 2) To maintain participants' attention. The testing vignettes were similar in structure and background settings, and without distractors, it could have been easy for participants to detect the study's aim. In addition, the repetitive scenes might have reduced participants' interest and attention, which lead to haphazard responses (Rungtusanatham et al., 2011).

The structure of the distractor vignettes was the same as the testing vignettes. However, their differences lie in the hostility towards the victim or the type of action performed by the perpetrator. In some distractors, instead of a hate action which causes the victim either negative consequences or nonexistent ones, a non-relevant or neutral action was involved (e.g. 'John briefly glanced at his watch to check the time and continued reading his book.'). In some others, the hostility shown was not related to religious hatred (the current interest of the study). Instead, the perpetrator's hostility was based on the victim's food or lifestyle preferences (i.e., hostility from a pro-animal activist against a meat-eater). An example of a distractor vignette read as follows (See Supplementary Information for Study Testing Materials): ' *In the* *supermarket, Emma saw Anna putting some salami and pork rib in her shopping trolley. Emma is vegan and strongly opposes the act of killing and eating animals. As Emma walked past Anna, she shouted at her, 'I wish you suffocate to death with your salami!'. As a consequence, Anna stopped going to the supermarket alone. '*

2.2.3.3. Attention task.

An attention task appeared randomly throughout the experiment to ensure participants read both vignettes and instructions. The attention task involved a vignette-format text but included instructions to direct participants' ratings in the relevant questions. Participants had to rate the questions as instructed to pass the attention check. An example of an attention task read as follows: *'This is a test for us to make sure that you read all the scenarios very carefully. Please answer the following questions, rating to what extent you think Mary should be punished as "Not at all (0)"; while rating how harmful Mary's action was as "Very Much (6)"'*.

2.2.4. Procedure.

We conducted the study using the Qualtrics software (www.qualtrics.com). After eliciting informed consent at the start, all vignettes, distractors, and attention tasks were presented randomly to the participants.

The study measured three dependent variables: Appropriate punishment, the likelihood of denouncing perpetrators to competent authorities, and the level of harm inflicted on the victim. Participants were presented with relevant questions and asked to rate their responses on a 7-point, forced-choice, continuous Likert scale, ranging from 0 (not at all, nothing at all or very unlikely) to 6 (very much, very likely, or extremely). With the forced-choice scale, participants had to process each question and provide a response (Allen, 2017) to proceed to the next one or the following vignette.

At the end of the experiment, demographic details (age, sex, educational background, degree of religiousness, social and political ideology and whether they have ever suffered a discriminatory experience) were collected. Finally, the last question is presented to gauge participants' awareness about legal sanctions against hate attacks in their country of residence (i.e. Are hate incidents legally sanctioned in your country of residence?). Data collection was blinded since experimenters had no contact with the participants.

2.2.5. Pre-registered sampling plan, power analysis and exclusion criteria.

Participants were recruited through Prolific (https://prolific.co), an online testing platform considered a reliable source of data collection (Palan & Schitter, 2018), which provides the flexibility to expand the range of people and geographical areas that can be included in the sample (Rupert et al., 2017). This was especially useful since our study recruited only participants who currently live in the U.K. and have English as their first language.

For hypotheses H1a, H1b, H2a, and H2b, an a priori power analysis conducted using G-Power (Faul et al., 2007, 2009) software showed that a total of 1302 participants was needed to run a mixed within-between-subjects MANOVA with sufficient power (alpha at 0.05, power (1 – β) set to 0.95, effect size f (V) set to 0.1, two groups and two measures). Furthermore, since H1c and H2c assumed no significant differences in participants' scores, no minimal sample size was required to test those hypotheses.

We set recruitment parameters in Prolific to only choose individuals whose first language was English, ensuring that all participants fully understood the vignettes presented and avoiding possible misunderstanding of the vignettes and/or instructions given. In addition, incomplete and duplicate submissions were manually excluded too. Only submissions with a valid Prolific ID, which anonymously refers to a unique participant, were approved. Based on Prolific ID, we excluded duplicate submissions except for the initial one if it was complete and did not coincide with another submission by the same participant. Finally, participants who failed the attention check or took more than 15 minutes to finish with all the questions presented were excluded. Consequently, from a total of 1403 participants that were recruited, 1309 remained after applying the exclusion criteria mentioned above, complying with the sample size of the a priori power analysis.

2.2.6. Pre-registered analysis and results.

Data were pre-processed by applying the exclusion criteria mentioned above. Since our pilot study showed a correlation between dependent variables, as first step, we aimed to evaluate differences in participants' mean scores on the three dependent variables (appropriate punishment for the perpetrator, the likelihood of denouncing perpetrators to competent authorities and the level of harm inflicted on the victim), across levels of the independent factors, namely, action type (as a within-subjects factor: verbal and nonverbal hate actions) and consequence type (as a between-subjects factor: negative consequences and nonexistent consequences for the victim). All data analyses were performed using R (version 4.1.1).

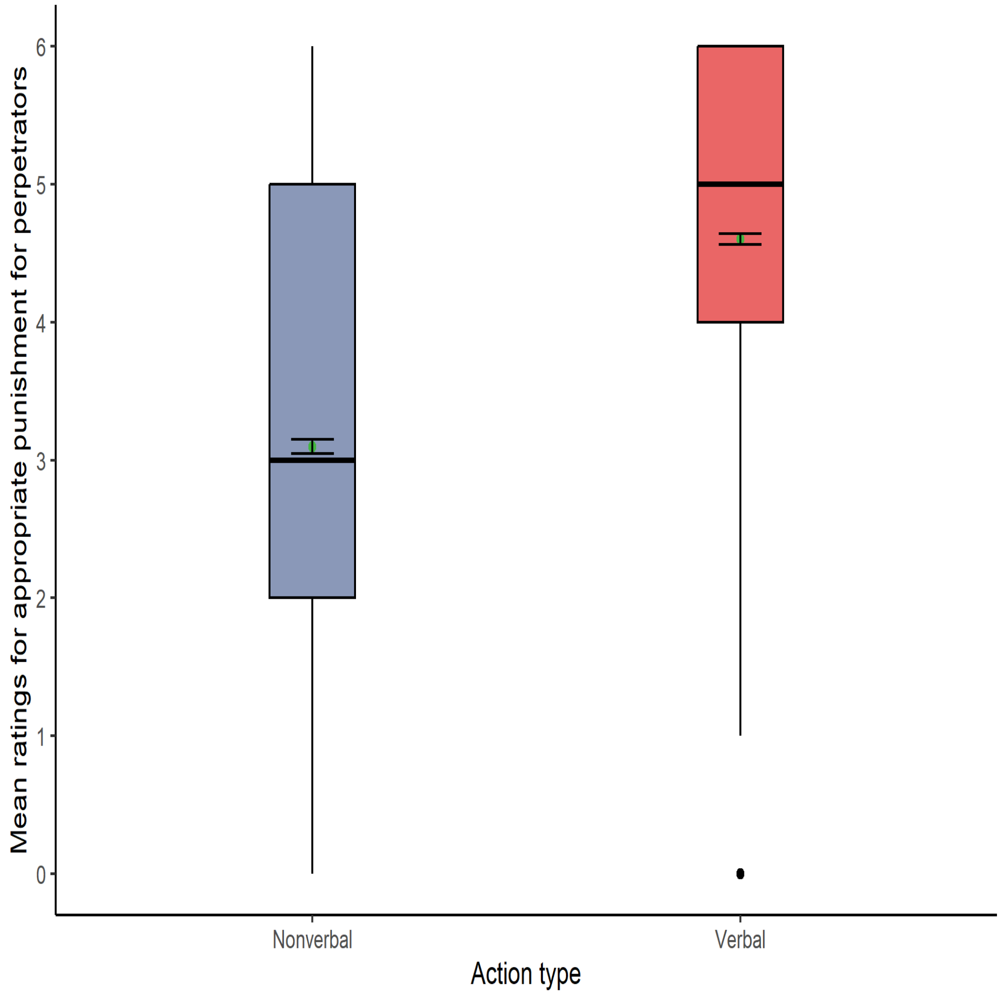
The following assumptions were assessed to run a mixed MANOVA: No multivariate outliers, homogeneity of variance and covariance, multivariate normality, Linearity and Multicollinearity. Multivariate outliers were tested by computing Mahalanobis distance for each observation, and eighteen (18) participants were identified as multivariate outliers ($p < .001$) and consequently removed according to the pre-registered protocol. Therefore the final sample size for the analysis was 1291 participants. The homogeneity of variance-covariance matrices was assessed via Box’s M test and the assumption of variance was violated (results showed $p< .05$). Therefore, as pre-registered, this violation was further investigated using Levene’s test for multiple independent variables. Again it showed violations ($p< .05$). Multivariate normality was checked using Mardia’s skewness and kurtosis test, and violations were found ($p < .05$). Linearity was assessed with scatterplots, and it was present. Finally, multicollinearity was tested for the three dependent variables and was not observed. No correlation was above $r = 0.90$ (Tabachnick & Fidell, 2012). Since the assumptions required for the MANOVA were violated, a Johansen's (Johansen, 1980) general formulation of Welch-James’s statistic with Approximate Degrees of Freedom (ADF), which is suitable for non-parametric mixed designs, was applied (Welch, 1951; Keselman, 2003; Villacorta, 2017) to evaluate differences in participants' mean scores on the three DV across the IVs. The results are reported in WJ format (df1, df2) for the Welch-James ADF statistic tests. The df1 and df2 are the approximate degrees of freedom for the numerator and denominator. Only results of $p < .05$ were considered statistically significant.

The Welch-James ADF test showed a significant main effect of the IV action type (WJ (3, 1041) = 351.69, df=1041, $p < .001$), a significant main effect of the IV consequence type (WJ (3, 1026) = 295.00, df = 1026, $p < .001$), and a significant interaction between action type and consequence type on the combined dependent variables (WJ (3, 1041) = 14.32, df=104, $p < .001$). Therefore, according to the pre-registered protocol, to further investigate those findings, a post hoc analysis was conducted for each dependent variable respectively.

Again, the following assumptions were previously tested: No outliers, normality, homogeneity of variance, and homogeneity of covariance. First, the Box Plot method was used to detect outliers, but none were found. Then, the data were analysed via histograms and Q-Q plots to test for normality, and the assumption was violated. Next, the homogeneity of variance was assessed via dot plots and Levine's test and it was violated (all results were $p < .05$). The homogeneity of covariance was evaluated via Box's M test, and it showed significant results for harm ($p < .001$) but not for punishing ($p = .105$) nor for reporting ($p = .992$). In addition, the assumption of sphericity was taken for granted since there were only 2 within-subjects levels. Since the assumptions required were violated, univariate testing for each of the three DV was conducted using the Holm-corrected Welch ADF test instead of using three mixed ANOVAs.

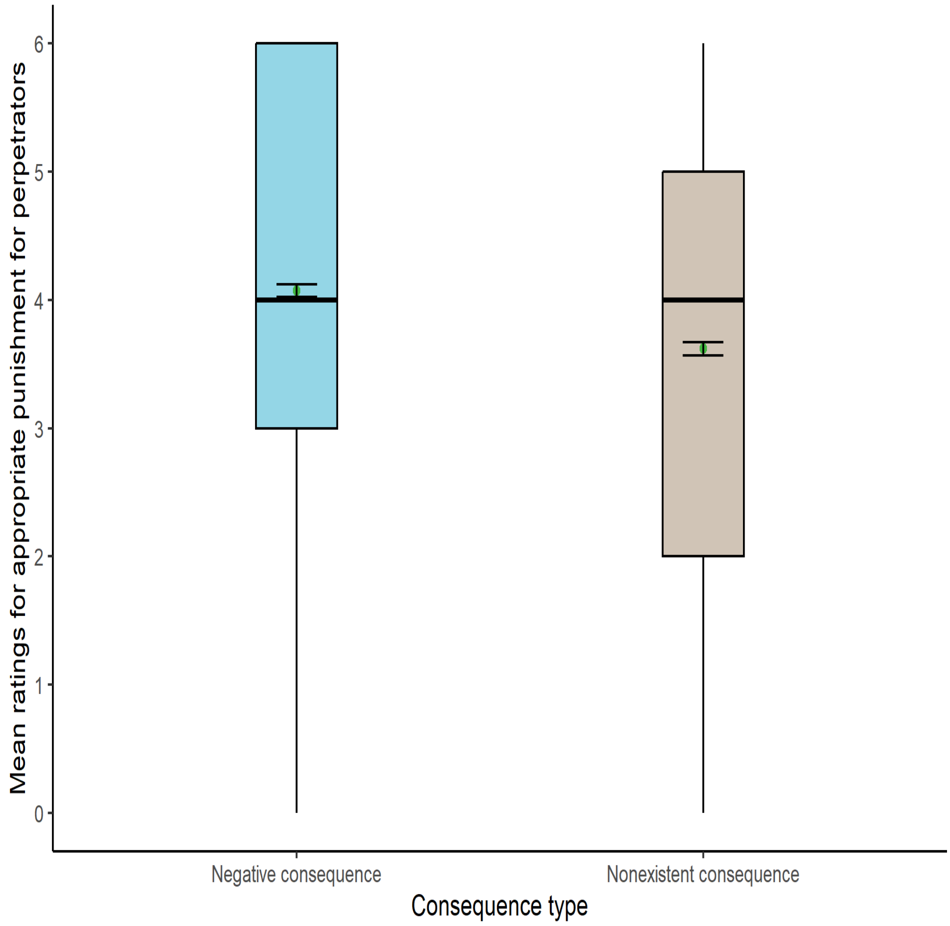
As predicted, the first Holm-corrected Welch-James ADF test for the dependent variable appropriate punishment showed a significant main effect of action type (Figure 1). However, **H1a** and **H2a** were not supported since the analysis showed higher mean scores for verbal (Mean = 4.60) than nonverbal hate actions (Mean = 3.10): WJ (1, 1287) = 938.99, p < .001.

Figure 1: Box plots show the appropriate punishment for each action type (N=1291).



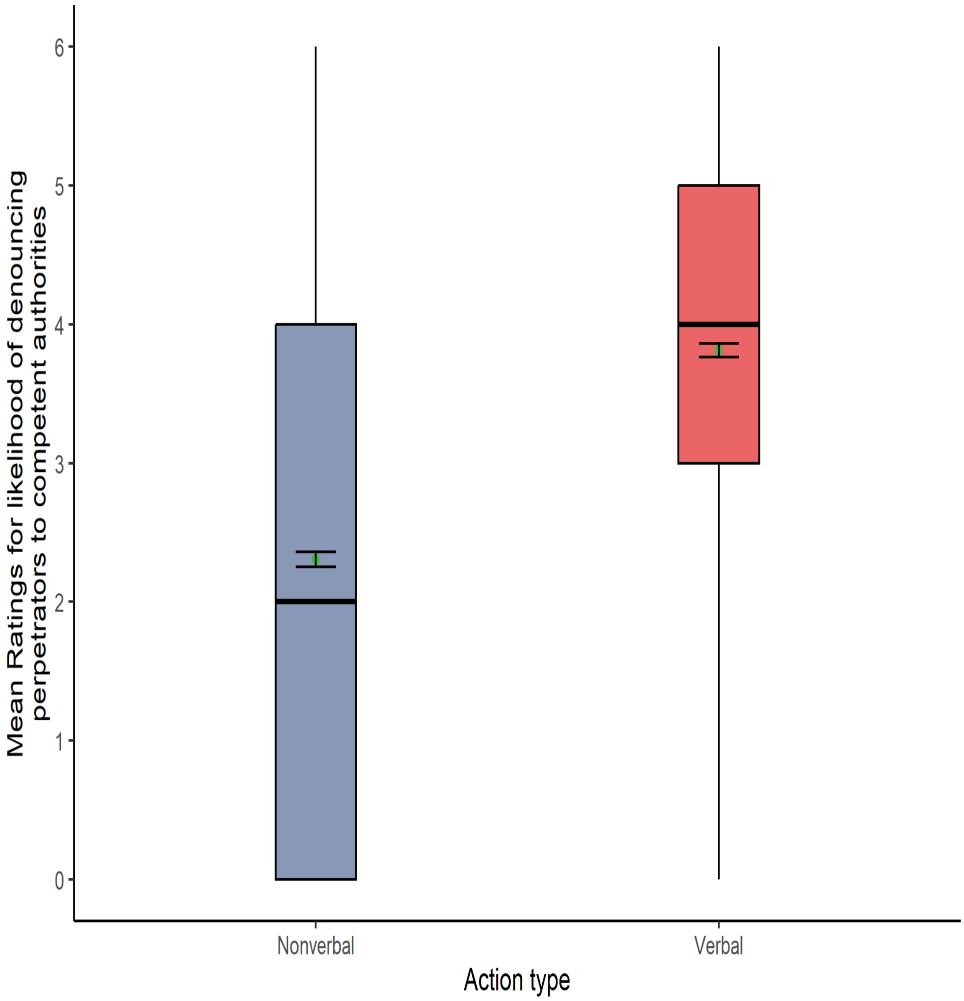
The analysis also showed a significant main effect of consequence type (Figure 2), with higher mean scores for negative (Mean = 4.07) than nonexistent consequences for the victim (Mean = 3.62): WJ (1, 1275) = 32.152, p < .001. No interactions between the two IVs were found.

Figure 2: Box plots show the appropriate punishment for each consequence type (N=1291).



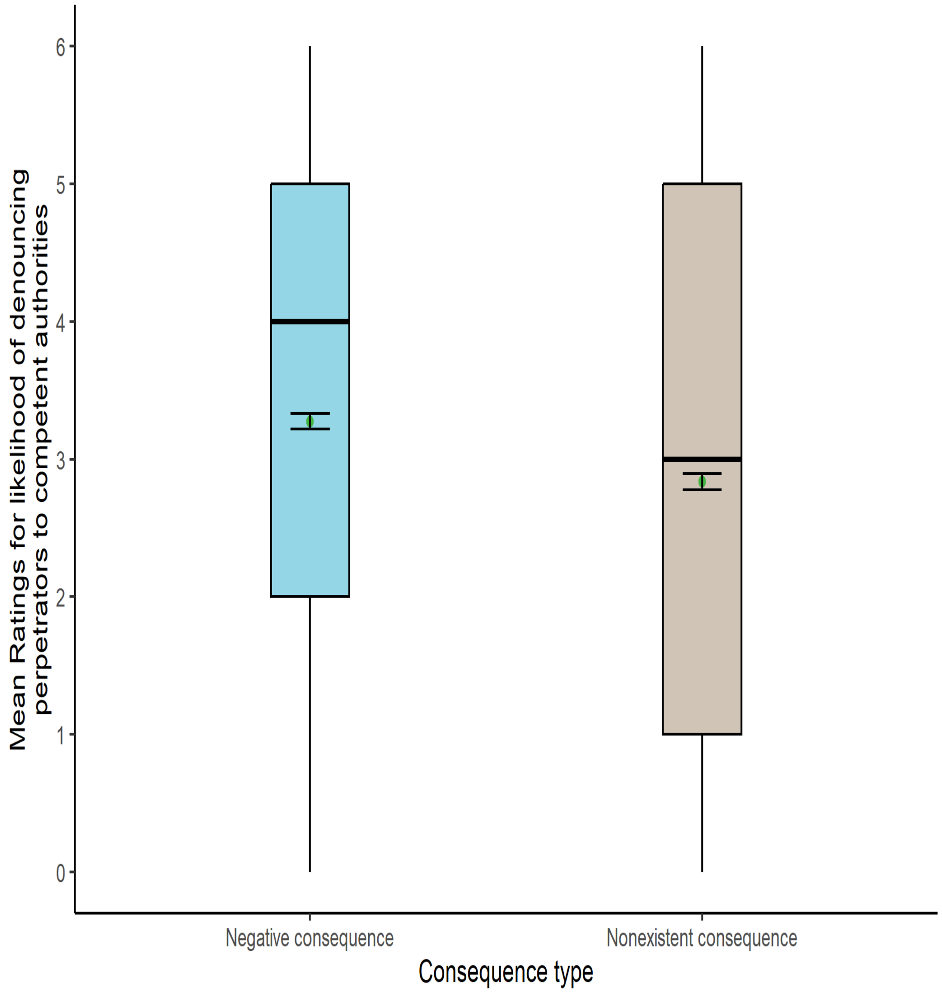
The second Holm-corrected Welch-James ADF test for the dependent variable likelihood of denouncing perpetrators to competent authorities showed a significant effect of action type (Figure 3), with higher mean scores for verbal (Mean = 3.81) than nonverbal actions (Mean = 2.31): WJ (1, 1286) = 797.64, p < .001. Therefore, **H1b** and **H2b** were unsupported.

Figure 3: Box plots show the likelihood of denouncing perpetrators for each action type (N=1291).



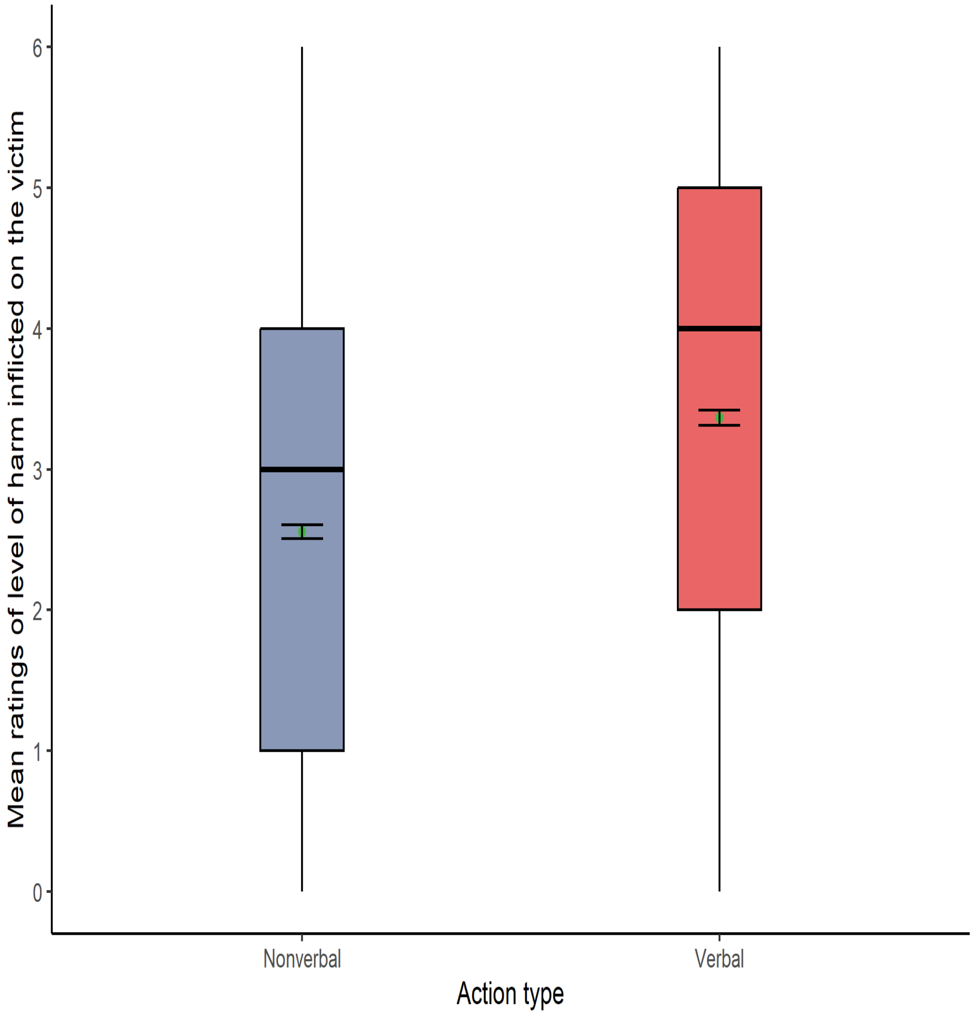
The analysis showed a significant main effect of consequence type also for this dependent variable (Figure 4), with higher mean scores for negative (Mean = 3.27) than nonexistent consequences for the victim (Mean = 2.84): WJ (1, 1286) = 22.96, p < .001. Again, no interactions between the two IVs were found.

Figure 4: Box plots show the likelihood of denouncing perpetrators for each consequence type (N=1291).



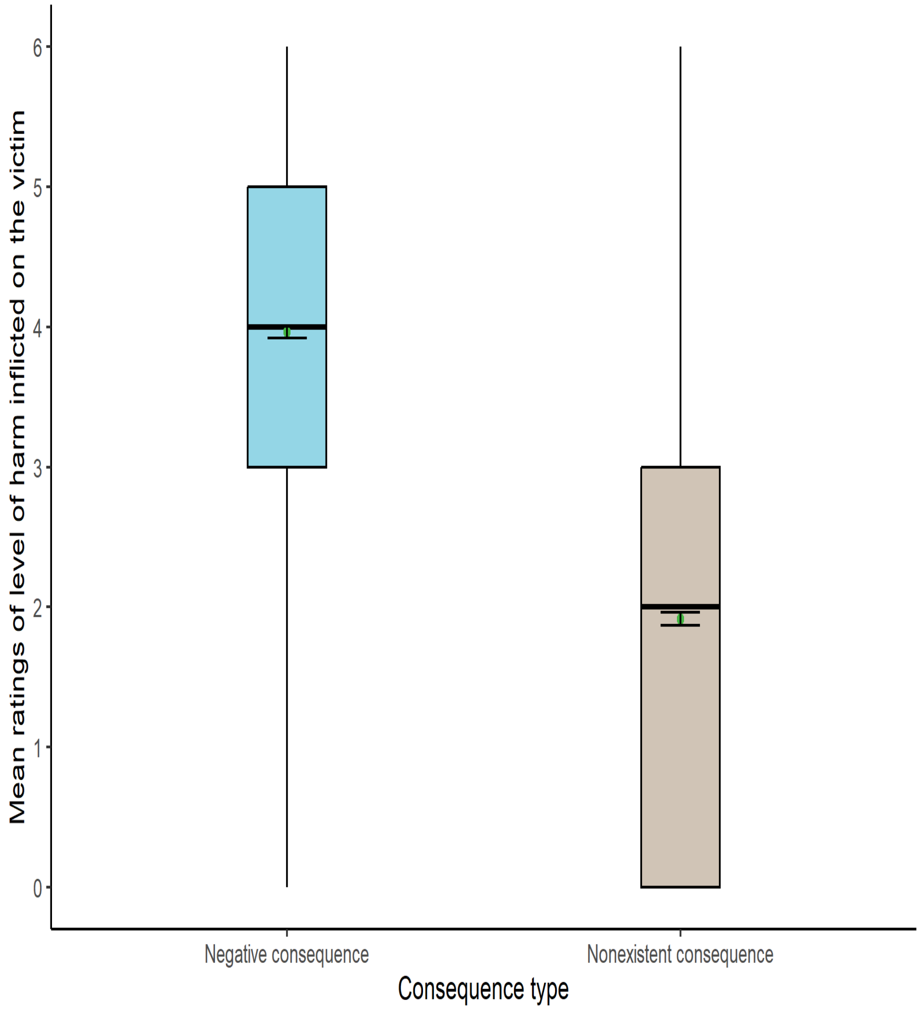
The third Welch-James ADF test for the dependent variable level of harm inflicted on the victim showed a significant effect of action type (Figure 5). Therefore, H1c and H2c were unsupported. Results consistently showed higher mean scores for verbal (Mean = 3.37) than nonverbal actions again (Mean = 2.55): WJ (1,1240) = 348.78, p < .001.

Figure 5: Box plots show the level of harm for each action type (N=1291).



In addition, results confirmed a significant main effect of consequence type (Figure 6), with higher mean scores for negative (Mean = 3.96) than nonexistent consequences for the victim (Mean = 1.91): WJ (1, 1227) = 763.80, p < .001.

Figure 6: Box plots show the level of harm for each consequence type (N=1291).



Finally, a significant interaction (Figure 7) between consequence type and action type was found: WJ(1, 1240)=30.83, p < .001. As pre-registered, since the data were ordinal, a Games Howell post hoc testing was conducted (instead of the Tukey method) to further analyse the interaction (Table 1 and 2). Results showed significant differences between all groups (p < .001).

Figure 7: Interaction plot between action type and consequence type in the level of harm inflicted on the victim (N=1291).

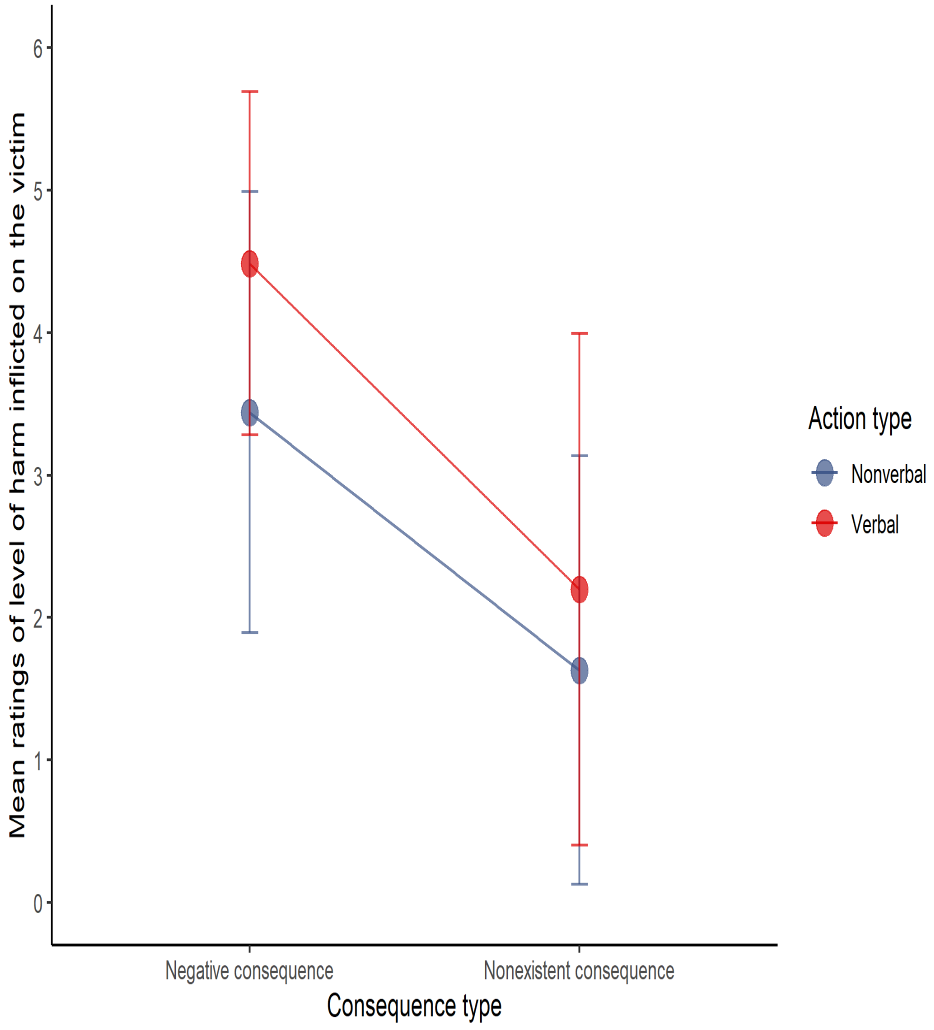


Table 1: Results of Games-Howell post-hoc test grouped by action type.

Action type	Group1	Group2	n1	n2	Estimate	CI	se	Statistic	df	p value
Non verbal	Non-existent consequence	Negative consequence	632	659	1.79	[1.62-1.93]	0.06	21.05	1288.70	<.001
Verbal	Non-existent consequence	Negative consequence	632	659	2.27	[2.10-2.44]	0.06	26.50	1098.37	<.001

Table 2: Results of Games-Howell post-hoc test grouped by consequence type.

Consequence type	Group1	Group2	n1	n2	Estimate	CI	se	Statistic	df	p value
Non-existent Consequence	Non-verbal	Verbal	632	632	0.57	[0.38-0.75]	0.06	6.11	1224.09	<.001
Negative consequence	Non-verbal	Verbal	659	659	1.05	[0.9-1.20]	0.05	13.76	1243.59	<.001

2.2.7. Exploratory analysis and results.

Consistent with our Stage 1 Registered Protocol, we ran an exploratory analysis to examine whether individual differences in sensitivity to verbal actions would predict differences in moral judgements of verbal hate attacks (See Supplementary Information for testing materials, analysis and results). We found positive, statistically significant, and medium correlations between ratings of sensitivity to verbal actions and ratings in deserved punishment, likelihood to denounce and the level of harm of verbal hate actions. Moreover, regression analyses showed that ratings of sensitivity to verbal actions significantly and positively predicted ratings in deserved punishment, likelihood to denounce and harmfulness of verbal hate actions. However, the questionnaire and the rating scale were not previously validated to test sensitivity to verbal actions. Therefore, we consider our results inconclusive and that a more rigorous scale and testing materials are needed to further investigate possible links.

In addition, to analyse the data collected further, we explored possible interactions with participants' demographics (age, sex, educational background, degree of religiousness, social and political ideology, previous discriminatory experiences and the awareness of legal consequences for hate incidents), which are reported as supplementary material.

2.2.8. Discussion

The high proportion of underreported hate speech attacks, the relatively few cases that end with sanctions, and the inconclusive evidence about moral attitudes towards these harmful incidents, made us wonder whether that leniency towards speech harm was ingrained in our moral dispositions, and consequently made us evaluate the harm caused by words as less harmful and worthy of punishment and denunciation than equivalent physical damage. Therefore, in an online

2x2 factorial experiment (N=1309) based on the contrastive vignette method, we tested action type and outcome aversion and its interaction in participants' evaluations of verbal and nonverbal hate incidents driven by the same hate intent and which create the same consequences for the victims.

Following the principle of action aversion, which explains that moral evaluation and punishment are determined not only by intentions and outcomes but also by associations with some intrinsic properties of the actions and people's aversion to them, we predicted that people would evaluate verbal and nonverbal hate-intended actions differently, being more lenient with verbal ones, even when both create the same consequences for the victims.

Expectedly, participants assessed hate actions with a negative outcome for the victim more severely than those that did not succeed in that purpose. In addition, they did evaluate verbal and nonverbal hate actions differently, irrespective their outcome. Therefore, our results provide evidence supporting both main effects, action and outcome aversion, in people's ratings of deserved punishment, denunciation and the level of harm inflicted on the victim, and a significant interaction of both effects in the last. However, contrary to our predictions, participants in our study consistently rated verbal hate actions as more worthy of punishment and denunciation and more harmful than nonverbal ones.

One possible explanation for these results could be that participants have found the hate intent more evident in hate speech than in other hate-motivated attacks. For example, the disdain and contempt towards the victims could be more explicit in verbal hate attacks, exacerbating participants' moral condemnation. Additionally, it could be the case that the harm in nonverbal attacks that do not reach the extreme of kicks and punches can be more plausibly denied by offenders and observers, providing them with moral wiggle room to remain inactive. Moreover, compared to a nonverbal hate attack, a verbal one has a host of related harms that go beyond the damage inflicted on the victim: that caused to bystanders, to other members of the targeted collective, and to society as a whole, which folk intuitions could capture.

Another possible explanation is that participants could have overestimated the degree to which a discriminatory comment would provoke their rejection. This would explain why the proportion of under-reported hate speech attacks continues to be striking despite participants self-report more severe evaluations of verbal hate actions in terms of punishment, denunciation and harmfulness. As pointed out by Kawakami et al. (2009), ordinary citizens usually fail to predict how they would feel and behave when faced with an act of racism. In their study, participants indicated that they would be very upset when witnessing such an incident, but they finally showed little emotional distress and responded indifferently. Therefore, although participants in our study reported they would be more severely against verbal hate actions, they may have wrongly anticipated their responses in real life. Further research is needed to test these possible explanations for a more severe response to harm caused by words.

2.2.9. Differences with pilot study results.

Two additional reasons could help to explain why we did not replicate the findings from our pilot study, in which participants assessed nonverbal hate attacks more severely than verbal ones while considering both as similarly harmful.

First, the pilot study tested a different version of scenarios A1 and A2. In them, the perpetrator, the victim, the perpetrator's hate intent, and the verbal and nonverbal actions were the same, but we avoided mentioning the consequences to participants.

Recently, Kneer & Skoczén (2023) showed that the folk concept of punishment is outcome-dependent. Furthermore, as they pointed out, prominent studies (Gino et al., 2009; 2008b) showed that learning about negative consequences can influence people's assessments of ethicality, to the point of assessing behaviours previously considered acceptable, as more unethical after being told about their consequences. In addition, as has been empirically demonstrated (Lench et al., 2015), in some cases, the simple consideration of alternative outcomes could alter participants' judgements. As a result, it is possible that introducing information about the consequences created for the victim affected participants' judgements, principally of deserved punishment and the likelihood of denouncing perpetrators. Therefore, in future work, we plan to specifically explore the effect of mentioning and not mentioning the action's negative consequences on participants before asking them to assess harmful verbal and nonverbal hate actions.

Second, another critical factor that could help to explain the variation between pilot and pre-registered study results is that both were tested during the pandemic of COVID19, but in totally different circumstances. On the one hand, the pilot study was conducted in November 2020, at the height of a global pandemic, with burdensome restrictions to prevent the spreading. And it tested a single root scenario, whose nonverbal version described a perpetrator spitting close to the victim's feet. Possibly, we underestimated the role of disgust in moral judgement in those circumstances. As some experts defend (Curtis et al., 2004), disgust is thought to have evolved as a biological mechanism that puts distance between us and anything that could potentially infect us. Therefore, presenting a scenario where someone spits close to the victim could have distressed participants overly, even more if they considered that the perpetrator might have removed his mask in doing so, an action that was expressly forbidden at that time.

On the other hand, the pre-registered study was tested in August 2022, when most people were fully vaccinated, and the rules of using masks and keeping their distance were lifted. This time the study also tested several verbal and nonverbal scenarios, one of the nonverbal described a perpetrator stopping the victim from sitting next to him in public transport, which at that time was normalised as a measure that helps to control the spreading. As a result, it could be possible that the nonverbal action of spitting close to the victim's feet was assessed more severely in the pilot study than in the pre-registered one. Again, additional investigations are needed to test these explanations for a more severe response to harm caused by words.

Finally and in addition to the above, our results replicate, for verbal (speech) harm, the moral luck phenomenon tested by Kneer & Skoczén (2023) in a recent study with implications for social psychology and moral theories. This phenomenon makes people assessing potential harm as more likely when it does come to pass than when it does not and, therefore, they judge unlucky perpetrators (whose actions ended in adverse outcomes) more severely than lucky ones (whose actions were neutral or ended harmlessly due to an external event). In our study, participants were randomly distributed into 2 groups: Those in group A tested two experimental trials (verbal and nonverbal), ending with the same negative consequences.

Participants in group B tested the same experimental trials. Still, this time the victim luckily did not suffer the expected consequences due to an external event (e.g., the victim was deaf and could not hear the hateful remark or was distracted and did not see the perpetrator's reaction). Our results showed that participants judged the unlucky perpetrators more severely than the lucky ones in terms of punishment, denunciation and the level of harm inflicted on the victim.

2.2.10. Data availability and supplementary materials.

The data sets generated and analysed for this study and all supplementary and testing materials are available through the Open Science Framework:

https://osf.io/wbasx/?view_only=6d80e55117704031bb6de41a4c99ef4f

2.2.11. Code availability.

Custom code that supports the findings of this study is also available through the Open Science Framework:

https://osf.io/wbasx/?view_only=6d80e55117704031bb6de41a4c99ef4f

2.2.12. Supplementary materials

a. Main study supplementary testing materials (Experimental trials)

Scenario setting A: Bus

Nonverbal A1 with negative consequences

Bilal, a Muslim man, was coming home by bus after prayers at the mosque. Peter was another passenger on the bus who had often shown an intolerance of Muslims. He had never met Bilal but saw him leaving the mosque before getting on the bus. This bus was the only public transport between the mosque and Bilal's apartment, and the walk would otherwise take him an hour. When the bus got to his stop, Bilal had to walk past Peter. Peter looked Bilal straight in the eyes and spat on the floor next to him. As a consequence, Bilal stopped using that bus line for a month.

Verbal A2 with negative consequences

Bilal, a Muslim man, was coming home by bus after prayers at the mosque. Peter was another passenger on the bus who had often shown an intolerance of Muslims. He had never met Bilal but saw him leaving the mosque before getting on the bus. This bus was the only public transport between the mosque and Bilal's apartment, and the walk would otherwise take him an hour. When the bus got to his stop, Bilal had to walk past Peter. Peter yelled at him, "Go home! Stop Islamization of our country!" As a consequence, Bilal stopped using that bus line for a month.

Nonverbal A3 with nonexistent consequences

Bilal, a Muslim man, was coming home by bus after prayers at the mosque. Peter was another passenger on the bus who had often shown an intolerance of Muslims. He had never met Bilal but saw him leaving the mosque before getting on the bus. This bus was the only public transport between the mosque and Bilal's apartment, and the walk would otherwise take him an hour. When the bus got to his stop, Bilal had to walk past Peter. Peter looked Bilal straight in the eyes and spat on the floor next to him. Bilal had very poor eyesight, so he did not notice Peter's reaction.

Verbal A4 with nonexistent consequences

Bilal, a Muslim man, was coming home by bus after prayers at the mosque. Peter was another passenger on the bus who had often shown an intolerance of Muslims. He had never met Bilal but saw him leaving the mosque before getting on the bus. This bus was the only public transport between the mosque and Bilal's apartment, and the walk would otherwise take him an hour. When the bus got to his stop, Bilal had to walk past Peter. Peter yelled at him, "Go home! Stop Islamization of our country!" Bilal used his wireless headphones and listened to loud music, so he could not hear a word coming from Peter.

Scenario setting B: Train

Nonverbal B1 with negative consequences

Ali, an elderly Muslim man, was returning home by train after a religious festival. Mark was another passenger sitting on the train who had always despised Muslim people. He had never met Ali but saw that he was wearing a Tarbush (A traditional-Muslim red hat). The seat next to Mark was the only one free in the waggon. Ali saw it and was heading to take it, but Mark scowling, put his backpack promptly on the free seat, stopping Ali from sitting next to him. As a consequence, Ali stood the whole journey.

Verbal B2 with negative consequences

Ali, an elderly Muslim man, was returning home by train after a religious festival. Mark was another passenger sitting on the train who had always despised Muslim people. He had never met Ali but saw that he was wearing a Tarbush (A traditional-Muslim red hat). The seat next to Mark was the only one free in the waggon. Ali saw it and was heading to take it, but Mark scowling, yelled at him, "Go where you came from! You are making our country sick!" As a consequence, Ali stood the whole journey.

Nonverbal B3 with nonexistent consequences

Ali, an elderly Muslim man, was returning home by train after a religious festival. Mark was another passenger sitting on the train who had always despised Muslim people. He had never met Ali but saw that he was wearing a Tarbush (A traditional-Muslim red hat). The seat next to Mark was the only one free in the waggon. Ali saw it and was heading to take it, but Mark scowling, put his backpack promptly on the free seat, stopping Ali from sitting next to him. Suddenly, a closer seat was left free, so Ali took it without even noticing Mark's reaction.

Verbal B4 with nonexistent consequences

Ali, an elderly Muslim man, was returning home by train after a religious festival. Mark was another passenger sitting on the train who had always despised Muslim people. He had never met Ali but saw that he was wearing a Tarbush (A traditional-Muslim red hat). The seat next to Mark was the only one free in the waggon. Ali saw it and was heading to take it, but Mark yelled at him, "Go where you came from! You are making our country sick!" Ali was attentively listening to his favourite audiobook, so he could not hear a word coming from Mark.

Scenario setting C: Supermarket

Nonverbal C1 with negative consequences

Hamza, a Muslim man, was shopping at his local supermarket. Harry was another client at the supermarket who had often shown an immense disdain for Muslim people. He had never met Hamza but saw him asking for halal meat (meat that meets requirements that Muslims consider to make it suitable for consumption). Harry realized that there was only one package of Halal beef left. He looked at Hamza and, smiling derisively, put the package into his shopping cart. As a consequence, Hamza stopped going to that supermarket.

Verbal C2 with negative consequences

Hamza, a Muslim man, was shopping at his local supermarket. Harry was another client at the supermarket who had often shown an immense disdain for Muslim people. He had never met Hamza but saw him asking for halal meat (meat that meets requirements that Muslims consider to make it suitable for consumption). When heading to supermarket check-out, Hamza had to walk past Harry. Harry yelled at him, "Get out! Stop destroying our culture!" As a consequence, Hamza stopped going to that supermarket.

Nonverbal C3 with nonexistent consequences

Hamza, a Muslim man, was shopping at his local supermarket. Harry was another client at the supermarket who had often shown an immense disdain for Muslim people. He had never met Hamza but saw him asking for halal meat (meat that meets requirements that Muslims consider to make it suitable for consumption). Harry realized that there was only one package of Halal beef left. He looked at Hamza and, smiling derisively, put the package into his shopping cart. Hamza was texting on his cell phone and did not notice Harry's reaction.

Verbal C4 with nonexistent consequences

Hamza, a Muslim man, was shopping at his local supermarket. Harry was another client at the supermarket who had often shown an immense disdain for Muslim people. He had never met Hamza but saw him asking for halal meat (meat that meets requirements that Muslims consider to make it suitable for consumption). When heading to supermarket check-out, Hamza had to walk past Harry. Harry yelled at him, "Get out! Stop destroying our culture!" Hamza was loudly talking on his cell phone, so he could not hear a word coming from Harry.

Distractors:

Distractor 1 (Neutral): Neutral action in a religious hatred context

Ahmed, a Muslim man, was on a flight coming back home after a conference in Morocco. John was another passenger on the flight who had often shown an intolerance of Muslims. He had never met Ahmed but saw him leaving the prayer room at the airport before getting on the plane. When the plane landed and passengers were asked to leave it, Ahmed had to walk past John. John briefly glanced at his watch to check the time and continued reading his book.

Distractor 2 (Nonverbal BMW): Nonverbal hate action against environment polluters

Albert saw James stepping out of a brand-new BMW parked outside an offices' building on his way to work. As an environmental activist, Albert despises luxurious car manufacturers. He considers all car manufacturers should be banned from polluting the environment and worsening climate change. Seeing James leave the luxury car, Albert went forward and kicked the rear fender in front of him. As a consequence, James stopped using that car park for a month.

Distractor 3 (Verbal Vegan): Verbal hate action against carnivores

Emma saw Anna putting some salami and pork-rib in her shopping trolley in the Supermarket. Emma is vegan and strongly opposes the act of killing and eating animals. As Emma walked past Anna, she shouted at her, 'I wish you suffocate to death with your salami! '. As a consequence, Anna stopped going to a supermarket alone.

Attention check:

Daniel is at a restaurant that just opened in his neighbourhood. Because it is a Mexican restaurant, Daniel believes, please ignore the rest of the information for this scenario. This is a test for us to make sure that you are reading all the scenarios carefully. Please answer the following questions, rating to what extent do you think Mary should be punished as Not at all (0); while how harmful was Mary's action as Very Much (6). Again, this is only a test for this story. Continue with the following questions. Mary sits quietly and continues eating his meal.

b. **Questions battery**

1. To what extent should Peter's action be punished?

Not at all Moderately Very much

0———1———2———3———4———5———6

2. If you had witnessed Peter's action, how likely would you report it to competent authorities?

Very unlikely Not decided Very Likely

0———1———2———3———4———5———6

3. How much harm was caused to Bilal?

Nothing at all Moderate Extreme

0———1———2———3———4———5———6

c. Exploratory study supplementary testing materials:

Sensitivity to verbal actions (Verbal-action Aversion Test)

How upset would it make you to curse angrily at an old woman as part of a movie script?

Not at all Moderately Extremely

0———1———2———3———4———5———6

How upset would it make you to stab a fellow actor in the neck during a play using a stage knife with a retractable blade?

How anxious would it make you to give a speech in front of a large crowd?

How anxious would it make you to compete in a sports competition?

How happy would it make you to hear a colleague speaking well of you accidentally?

How angry would it make you to see your favourite sport-team lose a championship game?

How angry would it make you to see your best friend insulting an immigrant?

How upset would it make you to shoot a bullet at a consenting friend while he's behind a bulletproof glass?

How good would you feel after your boss congratulates you for doing a great job?

How embarrassed would it make you to make obscene gestures directed at your best friend behind their back?

How upset would it make you to see a friend yelling derogatory remarks at her mother on the phone while holding down the mute button?

How embarrassed would you feel if you realized you had toilet paper stuck to your shoe?

2.3. Conclusion.

Our results show that people are more averse towards hate-motivated verbal actions than physical ones when intention and consequences remain equal. Against our predictions, we demonstrate that "words could hurt more than actions" also holds for third-party observers, who self-reported they would punish hate discourses more than equivalent physical acts. These results extend and complement a range of findings showing third-party' moral evaluation and punishment heuristics are determined not only by intentions and consequences but also by associations with intrinsic actions' properties and people's aversion to them. We confirmed that no lower moral condemnation of verbal hate attacks is ingrained in ordinary people's moral dispositions, with implications for social psychology and moral theories.

Moreover, our results show that ordinary citizens are open to recognising the harm caused by words which is a solid basis for developing public policies that reinforce civic engagement against hate speech incidents. Additionally, these findings contribute to the discussion regarding the limits of the freedom of speech principle that set them in the actual harm caused to victims, reporting a folk intuitions' assessment of hate speech harms. In sum, our findings confirm that a better understanding of the psychological processes behind the moral condemnation of verbal harm, specifically hate speech, is needed to develop and implement efficient regulations and policies against such harmful discourses.

In this study, we deliberately avoided scenarios representing extreme verbal and nonverbal violence. In this context, our results show that ordinary citizens self-report a higher tendency to punish and denounce demeaning and discriminatory speech targeting members of minority or disfavoured identities than comparable nonverbal actions, which is relevant for legislative and policy efforts against hate speech. Nonetheless, future work could explore the limits of this comparison, contrasting folk intuitions about more extreme scenarios (e.g., comparing the use of slurs and death threats with episodes of punching, beating or kicking).

In addition, in the present study, we recruited only native English speakers from the UK since this country pioneered the implementation of hate speech regulations in Western Europe and currently invests the most economical and human resources in creating social awareness about verbal harm. However, hate speech is a growing concern which similarly affects many contemporary democracies around the world. Therefore, further research is needed to explore whether our results are replicable with different groups of participants (e.g. Americans or non-English speaking populations). Finally, we focused here on anti-Muslim hate speech because just over half of the hate-crime offences in the UK are recorded as racially or religiously aggravated. Again, further research could explore whether our results are replicable with different hate biases (e.g. hatred based on race, sexual orientation or physical and mental disabilities).

CHAPTER 2. PERCEIVING HATE SPEECH`S HARM: THE ILLUSION OF ORDINARY PEOPLE’S LENIENCY AGAINST HATE SPEECH

Chapter 3

Assessing bystander’s responses to hate speech: Collective opposition is more efficient than individual confrontation[2].

3.1. Introduction.

Is remaining silent when witnessing a hate speech attack harmful? Conversely, does speaking out against hate speech reduce the harm the attack creates? Given that this demeaning speech is harmful (Maitra & McGowan, 2012; Waldron, 2012; Walters, 2014b; Zapata & Deroy, 2023), most theoretical approaches to hate speech argue that silent bystanders could unintentionally support the aggressors (Langton, 2007, 2012, 2018a, 2018b; Maitra, 2004). This support could consist in letting perpetrators informally gain practical authority to express hateful derogatory statements (Langton, 2018a; Witek, 2013), normalising the verbal abuse of targeted victims (Ayala & Vasilyeva, 2016) and creating more stress and suffering for them and, by extension, society (Gelber & McNamara, 2016; Goldberg, 2010, 2020; Janson et al., 2009).

[2] **Important note: Important note:** The content of the present chapter reproduces in its entirety a preliminary version of a paper co-authored with Prof. Dr Ophelia Deroy, Dr. Justin Sulik, and Mr. Clemens von Wulffen, submitted for publication to the scientific journal Humanities & Social Sciences Communications (Springer Nature) under the title "Collective opposition to hate speech is more effective than individual confrontation". The final version, entitled "Bystanders' Collective Responses Set the Norm against Hate Speech", was finally published in the same journal on February 29th, 2024 (Zapata et al., 2024).

Under this assumption, intensive research has explored the impact of actively responding by showing opposition to hate speech (Álvarez-Benjumea, 2023; de Silva & Simpson, 2022; Gelber, 2012; Howard, 2021; Lepoutre, 2017, 2019); analysing the contextual determinants that favour or disfavour bystanders' intervention (Dessel et al., 2017; Dickter & Newton, 2013; Gibson et al., 2020; Gulker et al., 2013; Hornsey & Imani, 2004; Rovira et al., 2021; Wong et al., 2021); investigating the best practices on how and when to counter-argue hateful remarks (Fumagalli, 2021; Gagliardone et al., 2015, Lepoutre, 2017); and identifying which subjects are better placed to respond to hate speech (Ashburn-Nardo et al., 2014, 2020). Researchers have also shown that getting involved in counter speech might be extremely challenging and costly for individual bystanders and even more so for targets of hate speech, who are the actual victims of those attacks (Czopp et al., 2003; Dickter & Newton, 2013; Langton, 2018b; McGowan, 2018; Nielsen, 2012).

Yet the initial questions remain untested: Do ordinary citizens perceive hate speech incidents as more harmful when they occur in front of silent, passive bystanders? Do third-party observers consider bystanders who voice their opposition helpful in reducing the harm created by hate speech incidents? These are the two core questions our study aims to address.

A satisfactory response should also illuminate why or how bystanders' responses could reduce the harmful effects of hate speech. Here we hypothesised that people perceive the same attack as less harmful when it occurs in a place where showing opposition is the social norm in place. Besides showing that a normative social context significantly shapes individuals' attitudes towards racism (Blanchard et al., 1994; Monteith et al., 1996; Zitek & Hebl, 2007), researchers have shown that discrimination and its harms increase if society allows shared norms prohibiting discrimination to be eroded by whatever means (Barr et al., 2018). Then, we find it essential to answer whether people perceive the same attack as less harmful when it occurs in a place where showing opposition is the social norm.

Here, we take social norms to be unwritten rules and regularities that occur in a social context and create shared expectations within a group about how people should behave in certain situations (Bicchieri, 2016; House, 2018). They regulate social interactions in an informal and often subtle way by changing individuals' social expectations (Opp, 2001; Przepiorka et al., 2022). Some examples include tipping at a restaurant, choosing the proper way to greet a stranger, how we talk or eat, but also norms that support unpopular, inequitable, or dysfunctional social outcomes, such as the persistence of the gender pay gap, tolerance of hate speech, or female genital mutilation (Przepiorka et al., 2022). They are temporary and subject to change, as happened with the social rule of not smoking in enclosed spaces (Bicchieri & Mercier, 2014; Opp, 2002).

3.2. Studying responses to hate speech through visual vignettes

People's responses to demeaning and offensive language have been analysed mainly using written vignettes (e.g., Almagro et al., 2022; de Araujo et al., 2020; Swim & Hyers, 1999). This type of vignette consists of short, carefully constructed descriptions offering a systematic combination of characteristics of persons, objects or situations. It is widely used in social sciences to investigate respondents' beliefs, attitudes, or judgments (Atzmüller & Steiner, 2010). Their effectiveness has been demonstrated, especially in sensitive research topics such as abuse, trauma, stigma, social injustice, sexuality or mental health, where data quality benefited from participants distancing themselves from personal circumstances when answering surveys or questionnaires (Khanolainen & Semenova 2020).

However, written vignettes also face problems because they offer scarce contextual information due to their word limit, making it challenging to reflect the richness of real-life situations and contextual determinants crucial to understanding some problematic cases (Parkinson & Manstead, 1993). To address this, researchers have made use of artistic visual material, demonstrating that offering images in addition to written information allows participants to better understand the situations they evaluate (Holm et al., 2018; Khanolainen & Semenova, 2020), notably in sensitive topics such as bullying or verbal abuse. We followed that line of research and created a battery of cartoons as visually enhanced vignettes for our study.

Using cartoons allowed us to easily show participants many aspects of the incidents that otherwise would require extensive descriptions: specific features of the physical appearance and facial expressions of perpetrators, bystanders and victims; their body language, the physical distance between bystanders and the attack, the public nature of the space where the attack occurs, and most importantly, whether the bystanders present responded individually or collectively, following the majority or against it. For example, we could show the perpetrators' disdain and dislike for the victims or the defencelessness of the racialised victims through their facial expressions, and we could make it clear that all bystanders had the opportunity to react against the attack by locating them close to the incident and by directing their lines of sight to the attack. By including those features, we provided participants with relevant information about the incident's social context and, at the same time, reduced the scope of subjective interpretations, making the experiment less demanding and allowing participants to focus on the questions presented.

Researchers have shown that derogatory language is considered more or less permissible depending on whether it is used by someone that shares group membership with the target (Almagro et al., 2022; Henry et al., 2014). Thus, by standardising the appearance of perpetrators and bystanders as "white-skinned" people and victims as “dark-skinned", we make it explicit that victims and perpetrators belong to different ethnic groups. Similarly, using derogatory expressions tends to be considered more inappropriate when stated by a man rather than a woman (Fasoli et al., 2015); therefore, we included female and male perpetrators in the vignette battery. Using cartoons made it easier to take all those considerations into account.

As we are still far from a consensual definition of hate speech incidents (Anderson et al., 2022; Lepoutre et al., 2023), in this study, we characterise them as those performed by a perpetrator with a degrading and discriminatory intention towards a victim based on a particular personal characteristic (race or ethnic origin, religion, gender, physical or mental conditions, among others) of the latter (Zapata & Deroy, 2023). As our study focuses on racist hate-speech, the verbal expressions we presented to participants consist of generic, demeaning and discriminatory phrases targeting dark-skinned victims that send a symbolic message that they are unwelcome and unworthy of social respect (e.g. "You are making our country sick", "Go back home!"). We ran a pilot survey where we tested several common hate expressions. In the present study, we only included stimuli that were rated as similarly harmful. With all these measures, we aimed to minimise confounding variables that might otherwise interfere with the research focus of our study.

3.3. Experimental work.

3.3.1. Experiment 1.

3.3.1.1. Study Description.

In this first experiment, we investigated the effect of bystanders' silent response when facing a hate speech incident. We collected participants' responses regarding two dependent variables: (1) the incident's perceived level of harm and (2) the blame assigned to the perpetrator.

Regarding the latter, we concretely wondered whether people would consider silent bystanders to contribute to the damage caused by the perpetrator and blame them for their passive response. Following the distribution of responsibility principle (El Zein et al., 2019;

Keshmirian et al., 2022), we assumed that if participants blame silent bystanders, they would distribute the responsibility for the harmful outcome between them and the perpetrator and, therefore, assess the perpetrator as less blameworthy in scenarios with silent bystanders present.

In a within-subjects design, we tested 4 non-factorial experimental conditions. Table 1 lists these conditions, which we refer to as Scenario A, B, C and D. Scenarios A and B are individual scenarios and show incidents that occurred in front of a single bystander. Scenarios C and D are collective scenarios and show incidents that occurred in front of a group of three bystanders. This non-factorial design aimed to compare the effect of an individual remaining silent (Scenario A) to one voicing opposition (Scenario B), but also to test the impact of a bystander staying silent in collective settings, either following the majority reaction (Scenario C) or going against it (Scenario D).

Table 1. Experimental conditions tested in Experiment 1.

Bystanders' reactions	Type of scenario	Nº Silent bystanders	Nº Opposing bystanders
A 🤐	Individual	1	0
B 😮	Individual	0	1
C 🤐🤐🤐	Collective	3	0
D 🤐😮😮	Collective	1	2

Finally, as an exploratory question, we investigated whether people identify bystanders who witnessed a hate speech incident and who remain silent as implicitly supporting the perpetrator. To this end, we collected participants' responses regarding the number of perpetrator supporters they identified in each scenario.

We formulated the following hypotheses:

H1: An individual scenario with a silent bystander (Scenario A) will be assessed as more harmful than one with an opposing bystander (Scenario B).

H2: A collective scenario with more silent bystanders present (Scenario C with 3 silent bystanders) will be assessed as more harmful than one with fewer (Scenario D with 1 silent bystander).

H3: In the individual scenario with a silent bystander (Scenario A), the perpetrator will be assessed as less blameworthy than in that with an opposing bystander (Scenario B).

H4: In the collective scenario with more silent bystanders (Scenario C with 3 silent bystanders), the perpetrator will be less blameworthy than those with fewer (Scenario D with 1 silent bystander).

3.3.1.2. Participants.

We conducted a power calculation with G*Power software for a Friedman test (equivalent to a nonparametric repeated measures ANOVA) and a post hoc Wilcoxon Signed Rank test. Results showed that to detect an estimated small effect size of .15 with an alpha probability of

.05 and a power of .80, 62 participants were required for the Friedman and 290 for the Wilcoxon. We then recruited 353 British English-speaking participants through Cloud Research (Amazon Mechanical Turk). We recruited only British participants since the UK is a European leader in combating hate speech and creating social awareness about verbal harm. Therefore, we expected British citizens would be more aware of the effects of showing opposition or remaining silent when facing a hate speech incident (Zapata & Deroy, 2023).

Before the analysis, we excluded data from 24 participants: Six failed the attention check, 17 submitted incomplete surveys, and one submitted a duplicated data set. The final sample size included in the study was N=329 participants (114 female, 5 prefer not to say/non-binary).

3.3.1.3. Procedure.

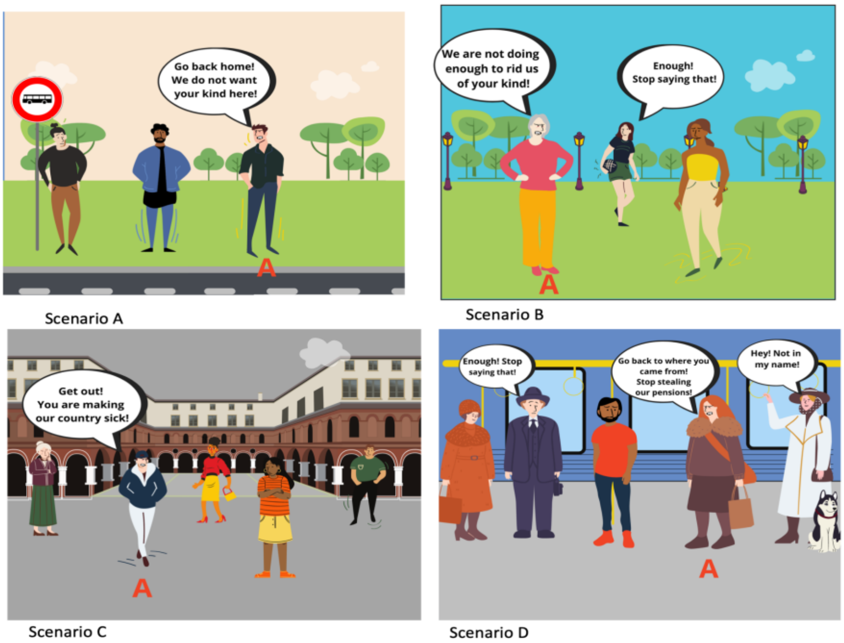
We conducted the study using the Qualtrics online platform (www.qualtrics.com). After providing informed consent, participants were shown four experimental scenarios and one attention check scenario (see below). Participants were asked to rate all scenarios see Table 1) regarding the incident's level of harm ("In your opinion, how harmful is the situation described above?") and the perpetrator's deserved blame ("To what extent should [perpetrator] *A* be blamed for the situation described above?"). The order of presentation of these DVs was randomised. Participants assessed all scenarios using a 7-point Likert scale ranging from 0 (Not at all) to 6 (Extremely).

Additionally, we presented participants with a question regarding the number of perpetrator supporters they identified in each scene ("How many [perpetrator] A's supporters do you identify?"). Participants responded using a forced-choice list that offers "zero", "one", and "two or more" as response options. All visual scenarios and their respective questions were shown in a randomised order. Participants finished the study by answering basic demographic questions. All participants who completed the survey and did not fail the attention check were paid 1.50 USD for a maximum of 8 minutes of work.

3.3.1.4. Testing Materials (Visual Vignettes).

We created a series of 16 colourful cartoons with a similar structure: All characters appear in a public space (A park, bus stop, street, or subway). A white-skinned perpetrator with an angry face yells a racist remark to a dark-skinned victim (e.g. "Go back home. We do not want your kind here!"), in front of one or three bystanders who witness the incident and either voice their opposition (e.g. "Enough! Stop saying that!") or remain silent. We aimed for consistency in the facial expressions, body language and skin colour of the perpetrators and victims. Perpetrators are angry-faced and show disdain and dislike for the victims; the victims appear alone and look intimidated or ashamed. The bystanders had a direct line of sight to the attack and were close to it. The scenarios were gender-balanced, with female and male perpetrators, victims and bystanders. Examples of the visual vignettes used are shown in Figure 1 (See Supplementary Information section for the complete battery of testing materials).

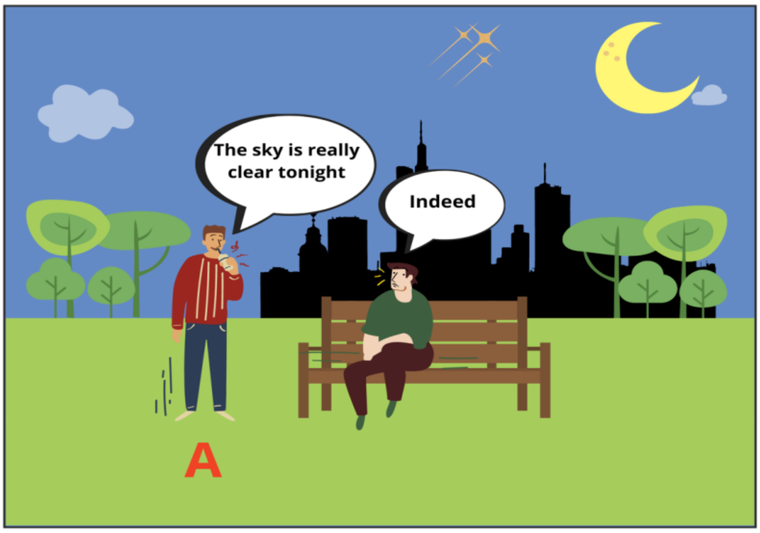
Figure 1. Example visual vignettes for each of the 4 experimental conditions (Scenarios A-D). The perpetrator is labelled as “A”.


Go back home! We do not want your kind here!
A
Scenario A
We are not doing enough to rid us of your kind!
Enough! Stop saying that!
A
Scenario B
Get out! You are making our country sick!
A
Scenario C
Enough! Stop saying that!
Go back to where you came from! Stop stealing our pensions!
Hey! Not in my name!
A
Scenario D

3.3.1.5. Attention-check Task.

An attention check appeared randomly throughout the experiment to ensure participants observed the experimental vignettes and read the questions (Fig. 2). The attention check had a vignette format. Still, it showed a friendly conversation between two people. Participants had to respond by assessing the incident as low in harm and the perpetrator as low in blame (below two on a 7-point Likert scale) to pass the attention check.

Figure 2. The attention check vignette (Experiment 1).



3.3.1.6. Analysis Strategy.

Data were pre-processed by excluding participants who failed the attention check. As we worked with Likert scales and ordinal data, we conducted a nonparametric Friedman test and a Wilcoxon signed rank test to analyse the differences in participants' median ratings (Sullivan & Artino, 2013) on the two dependent variables (*The incident's level of harm* and *the deserved blame for perpetrators*), across the four experimental conditions. All data analyses were performed in RStudio.

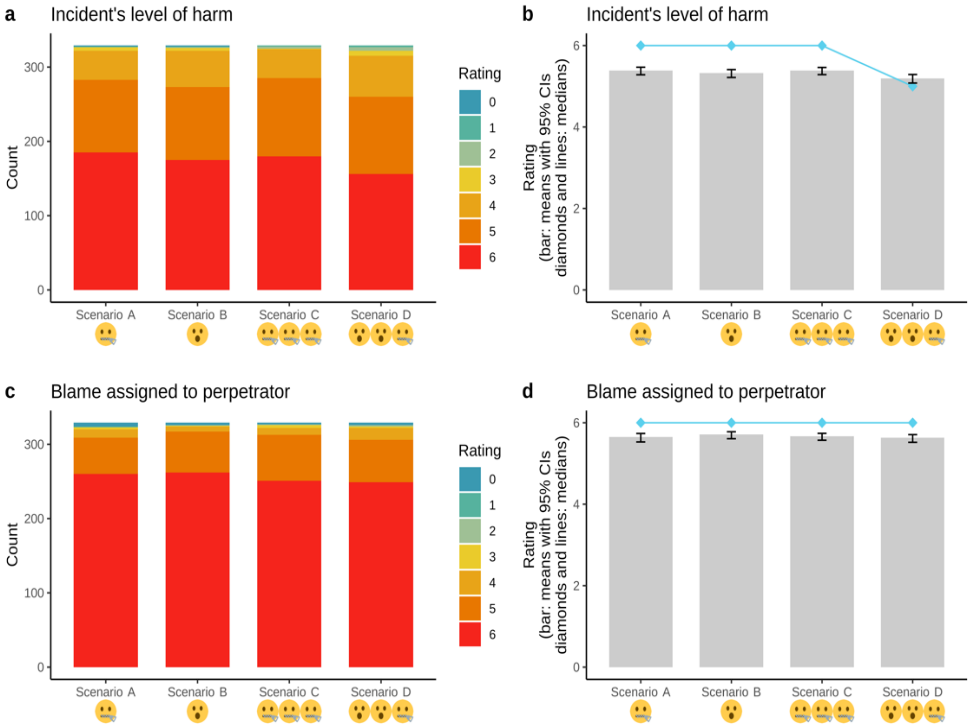
3.3.1.7. Results.

3.3.1.7.1. The incident's level of harm.

As expected, the results of a nonparametric Friedman test revealed significant differences in the ratings of the incident's level of harm between the experimental conditions ($\chi^2(3) = 27.06$, $p < .001$, Kendall’s W = .03 [.01, .05]). However, post hoc testing with Wilcoxon signed rank tests (and Holm-corrected *p*-values) revealed that there were no significant differences between scenarios A and B (both medians = 6, $r = .100$, $p_{adj.} = .477$), which rejects our first hypothesis (H1). In individual scenarios where a single bystander witnessed the attack, participants assessed the incident as similarly harmful, independently of whether the bystander showed opposition or remained silent. However, in collective scenarios with three bystanders present, participants assessed the scenario with more silent bystanders (Scenario C) as more harmful than that with fewer (Scenario D), confirming our second hypothesis (H2, Fig. 3a,b).

In addition, we found significant differences between scenario D (median rating=5) and all other scenarios (all other medians = 6; D vs A $r = .262$, $p_{adj.} < .001$; D vs B $r = .162$, $p_{adj.} = .015$; D vs C $r = .206$, $p_{adj.} < .001$). Thus, with more opposing bystanders present, Scenario D was assessed as the least harmful. Our results show that participants perceived bystander responses as beneficial only in collective settings.

Figure 3. (a) A stacked bar chart showing the distribution of ratings for the incident's level of harm, grouped by scenario; (b) grey bars show mean rating (and whiskers show 95% bootstrapped CIs) for the incident's level of harm; blue diamonds and lines show median responses; (c) A stacked bar chart showing the distribution of ratings for the blame assigned to perpetrators, grouped by scenario; and (d) grey bars show mean rating (and whiskers show 95% bootstrapped CIs) for the blame assigned to perpetrators, blue diamonds and lines show median responses.



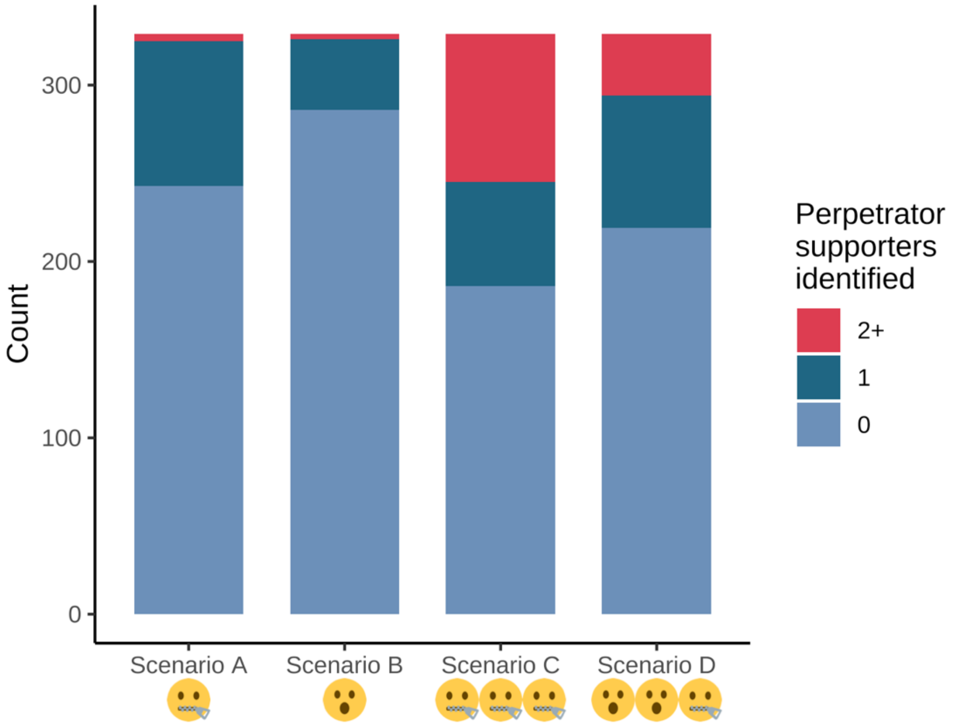
3.3.1.7.2. The perpetrators’ deserved blame.

Here, the analysis showed that the rating scores for the perpetrator's blameworthiness were not significantly different among scenarios. The results of a nonparametric Friedman test contradicted hypotheses 3 and 4 and revealed that median ratings for blame were not significantly different ($\chi^2(3) = 5.71$, $p = .127$, Kendall’s W = .006 [.001, .02]). Participants blamed perpetrators similarly, disregarding whether they attacked the victim in front of silent or opposing bystanders and whether the attack occurred in individual or collective settings (Fig. 3c, d).

3.3.1.8. Exploratory analysis.

We explored whether people tend to identify silent bystanders as supporting the perpetrator. To do so, we conducted a Cumulative Link mixed model regression analysis to test whether the number of bystanders present predicts the number of perpetrator supporters identified. We found that the number of silent bystanders was a significant positive predictor of perpetrator supporters identified (b = 0.60 [0.43, 0.77], SE = 0.08, $t = 7.02$, $p < .001$). Scenario C, with three silent bystanders present, was rated as having the highest number of perpetrator supporters (Fig. 4).

Figure 4. Graph showing the perpetrator supporters identified grouped by scenarios.



However, our design did not address whether—when counting perpetrator supporters—participants considered only silent bystanders or considered the silent victim too. Therefore, we address the issue of silent vs opposing responses in a more controlled manner in Study 2.

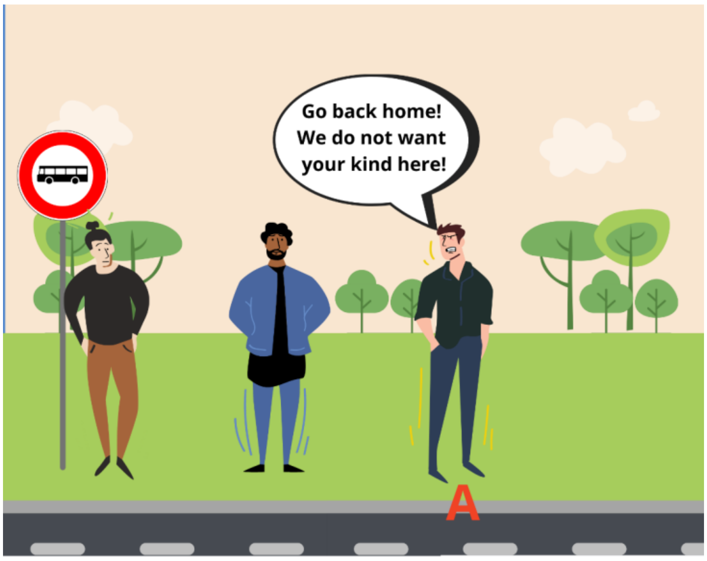
3.3.1.9. Discussion.

Experiment 1 showed that bystanders' reactions affected the perception of the harm caused by a hate speech attack only in collective settings when other bystanders are shown. This might suggest that when we do not offer participants enough elements to intuit the social norm against hate speech (by showing them how other bystanders react), they evaluate both hate incidents as similarly harmful, independently of whether the bystander present responded by remaining silent or showing opposition. Additionally, our results suggested that people evaluate scenarios where a group of bystanders voiced their opposition as less harmful than those where a group remained silent (Fig. 3a, b).

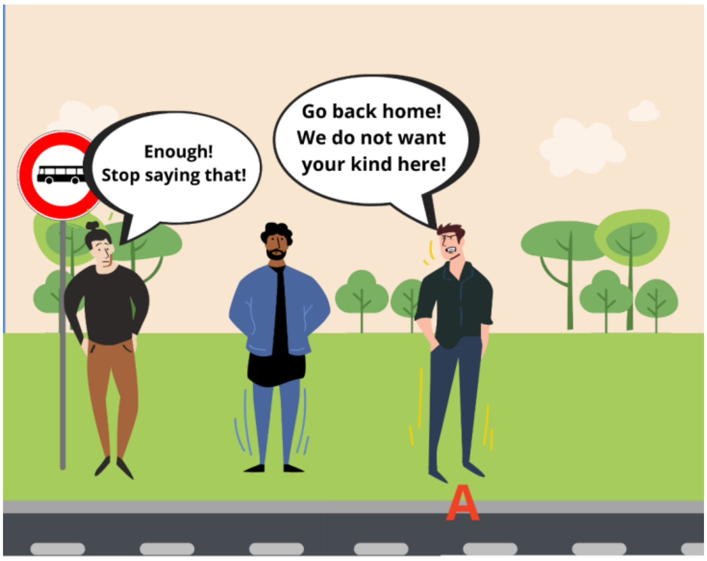
However, it remains unclear under precisely which conditions the perception of harm was affected by bystanders' opposing a hate speech attack in collective settings: Does opposing hate speech against the social norm—when the majority remains silent—affect the perceived damage differently than opposing it following the majority? We ran a second experiment to answer these questions.

3.3.1.10 Supplementary materials

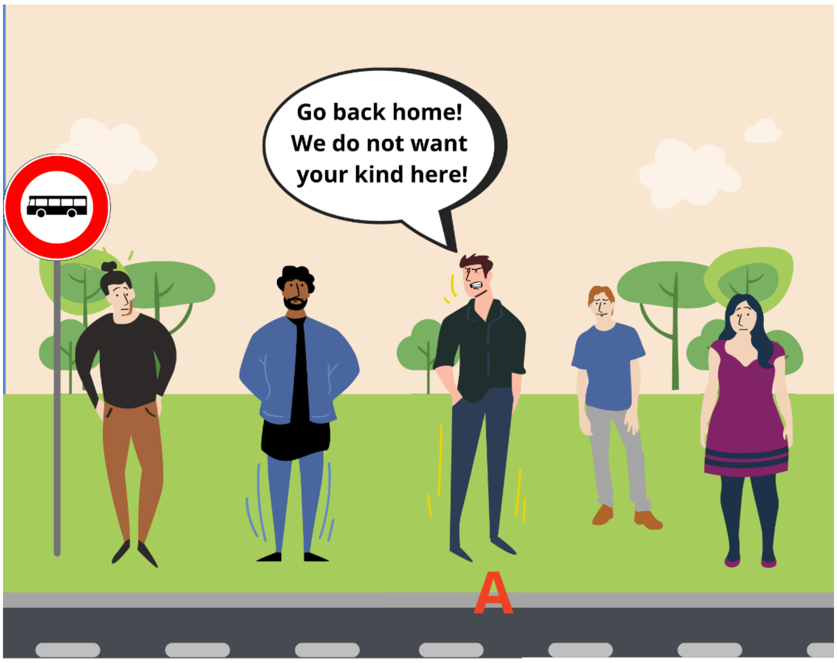
Visual vignettes



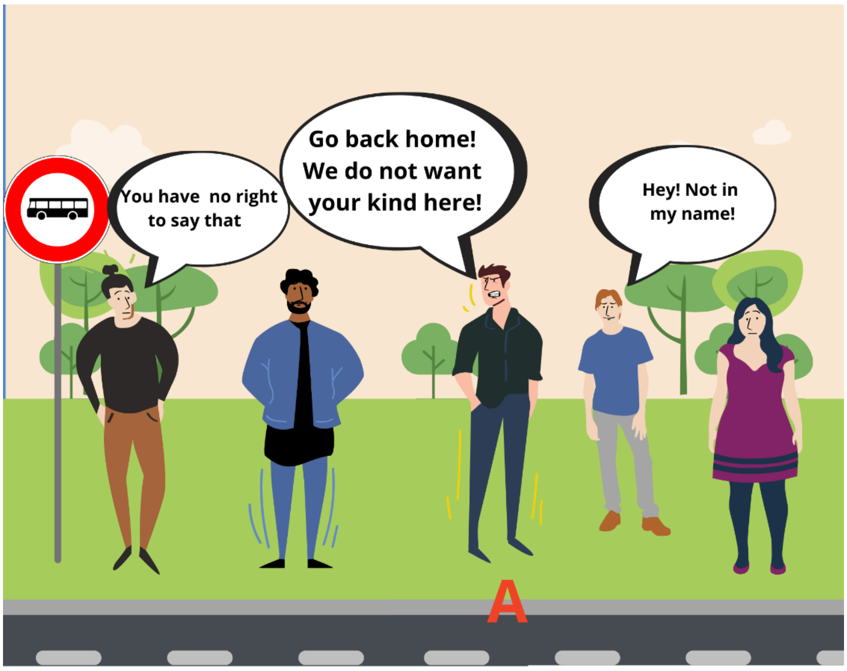
Individual Scenario with 1 silent bystander (Bus stop)


Enough! Stop saying that!
Go back home! We do not want your kind here!
A

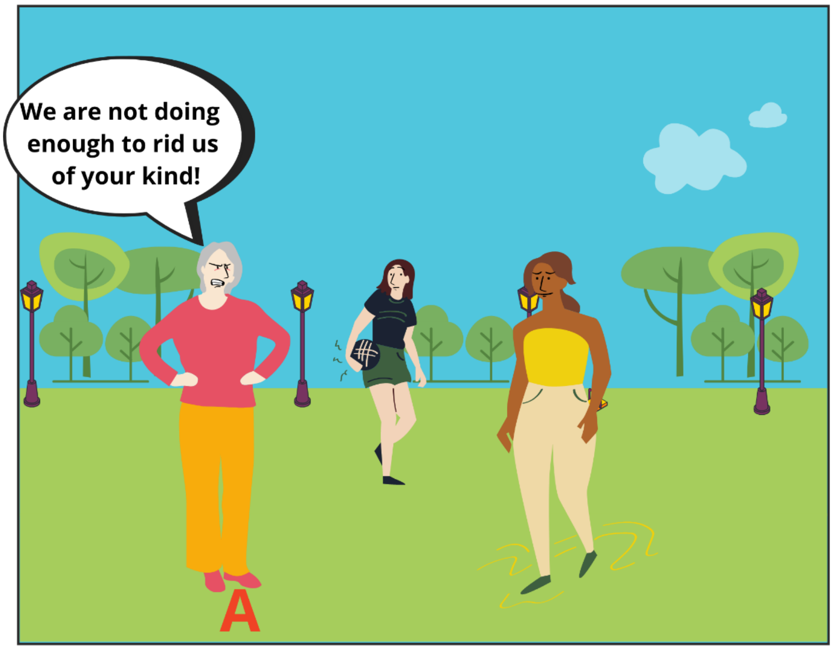
Individual Scenario with 1 opposing bystander (Bus stop)



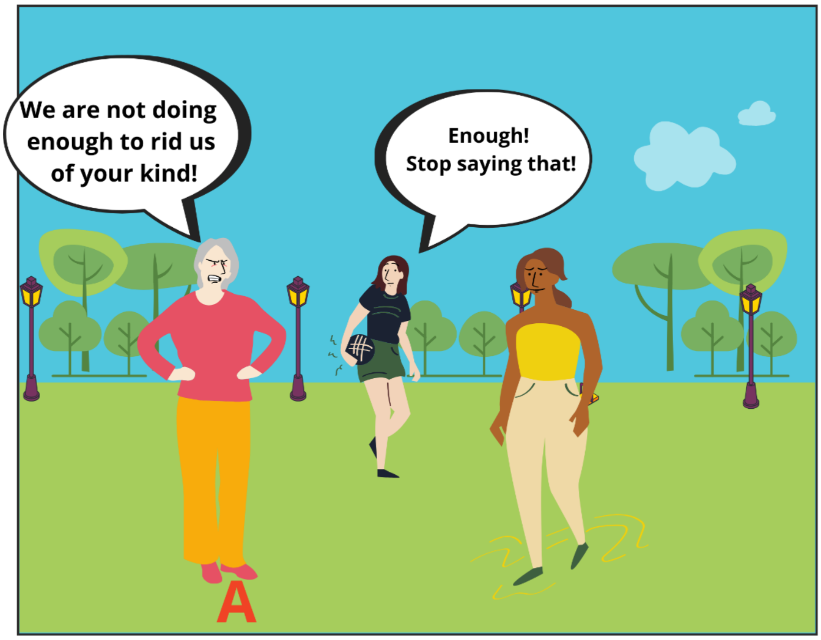
Collective Scenario with 3 silent bystanders (Bus stop)



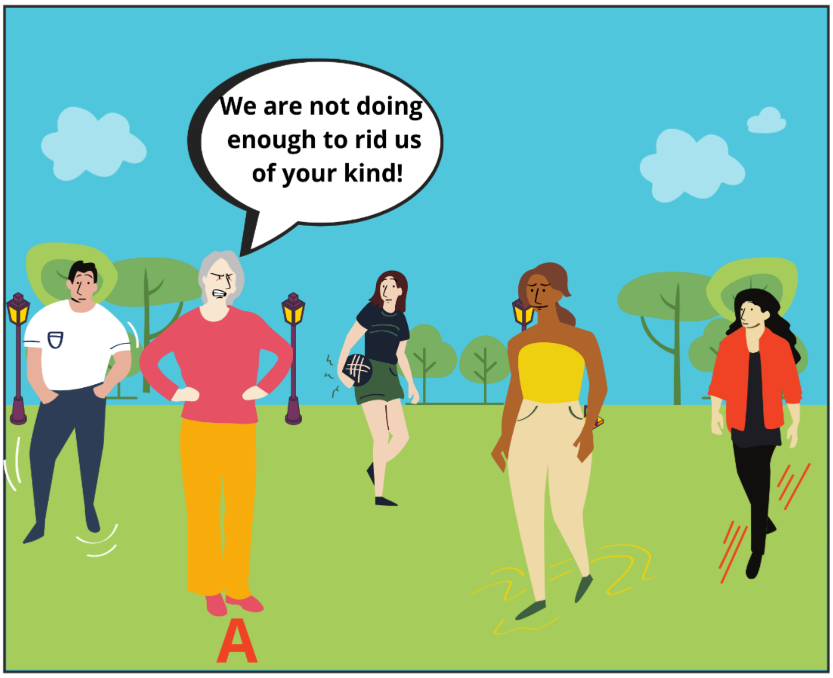
Collective Scenario with 1 silent bystander and 2 opposing (Bus stop)



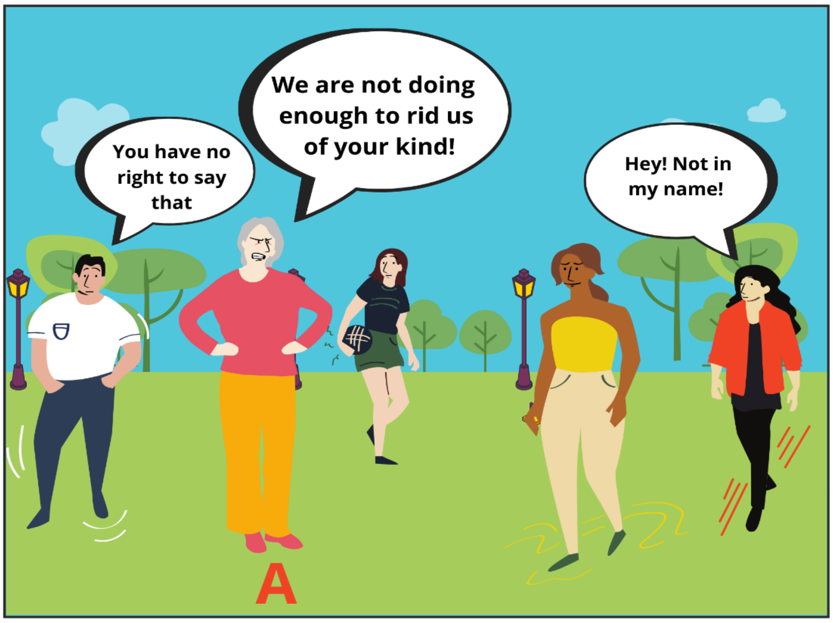
Individual Scenario with 1 silent bystander (Park)


We are not doing enough to rid us of your kind!
Enough! Stop saying that!
A

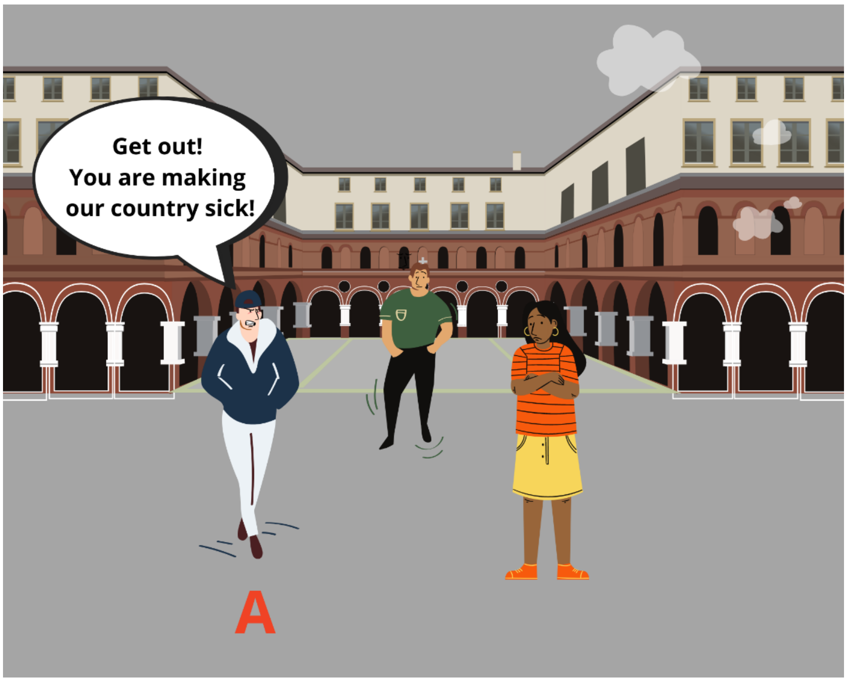
Individual Scenario with 1 opposing bystander (Park)



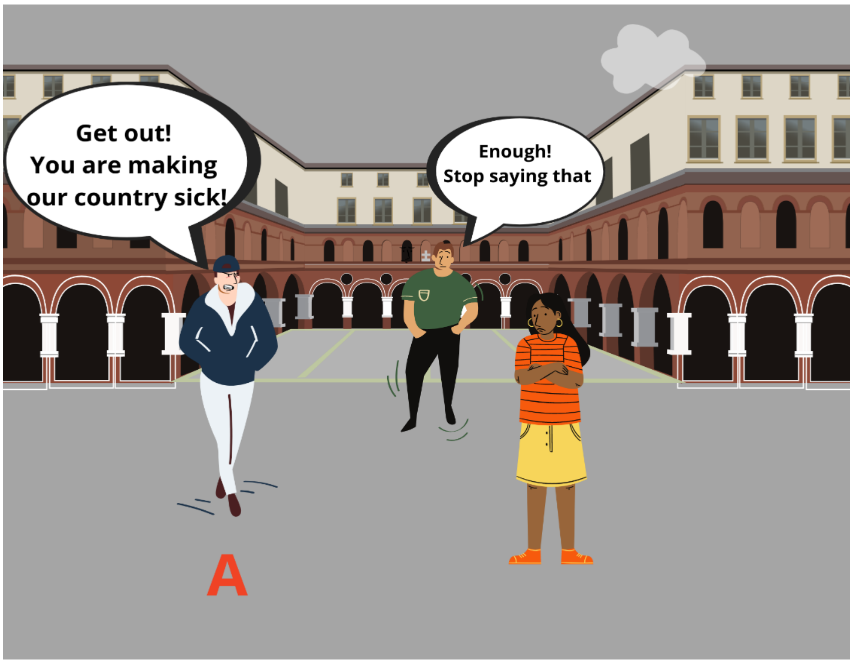
Collective Scenario with 3 silent bystanders (Park)


You have no right to say that
We are not doing enough to rid us of your kind!
Hey! Not in my name!
A

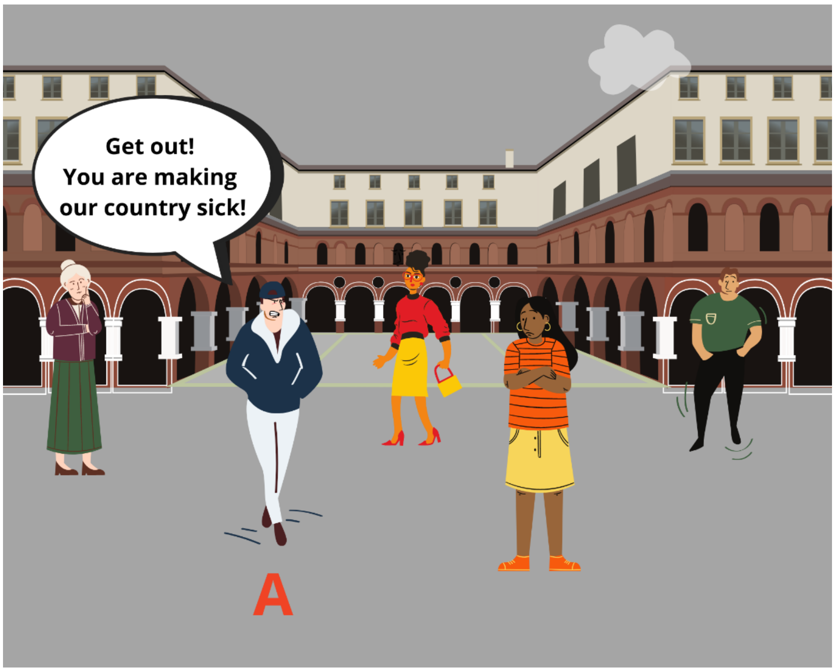
Collective Scenario with 1 opposing bystander (Park)



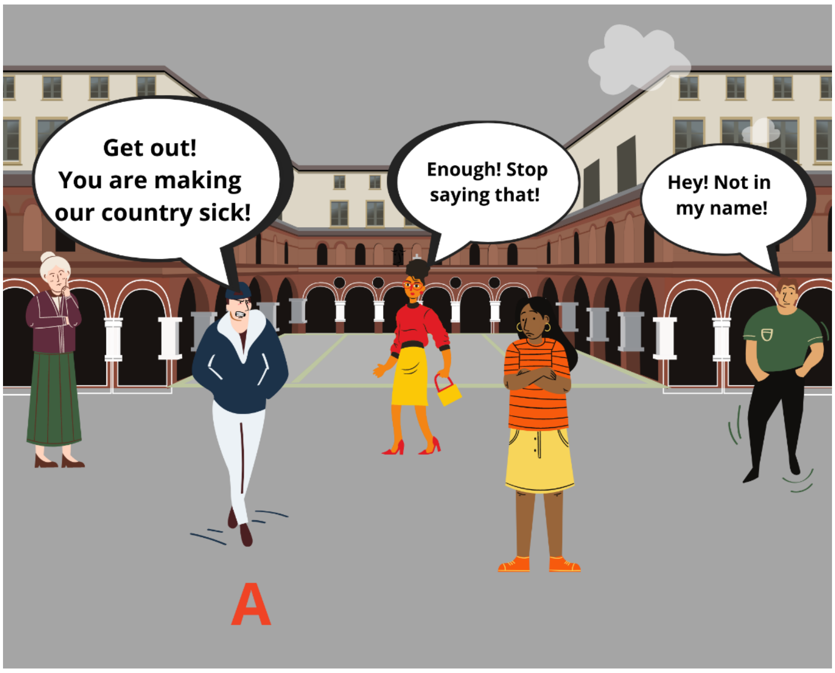
Individual Scenario with 1 silent bystander (Street)



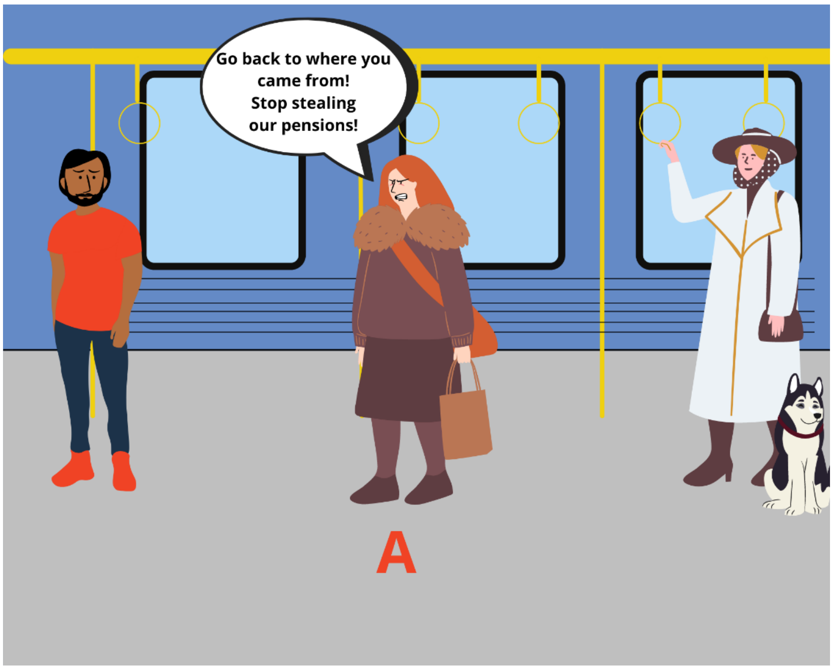
Individual Scenario with 1 opposing bystander (Street)



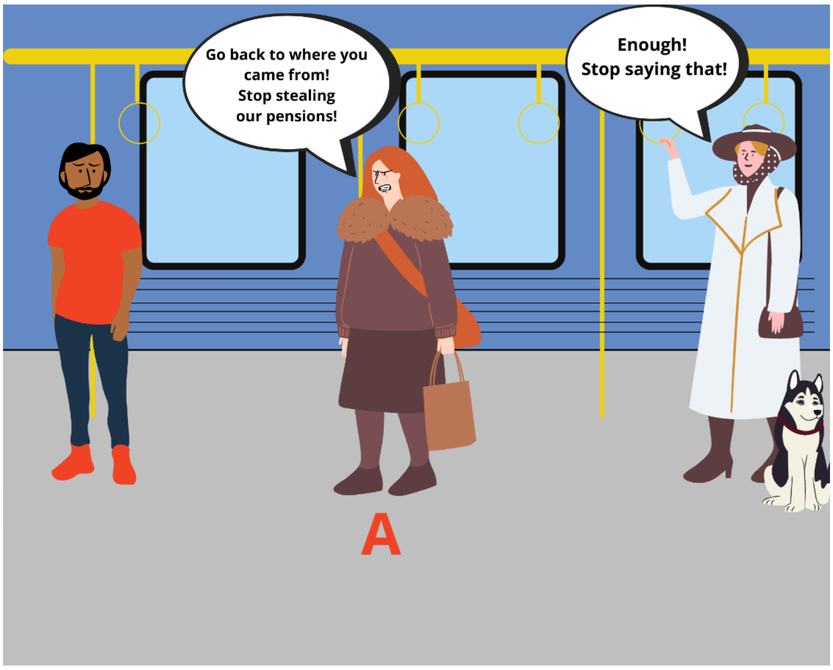
Collective Scenario with 3 silent bystanders (Street)



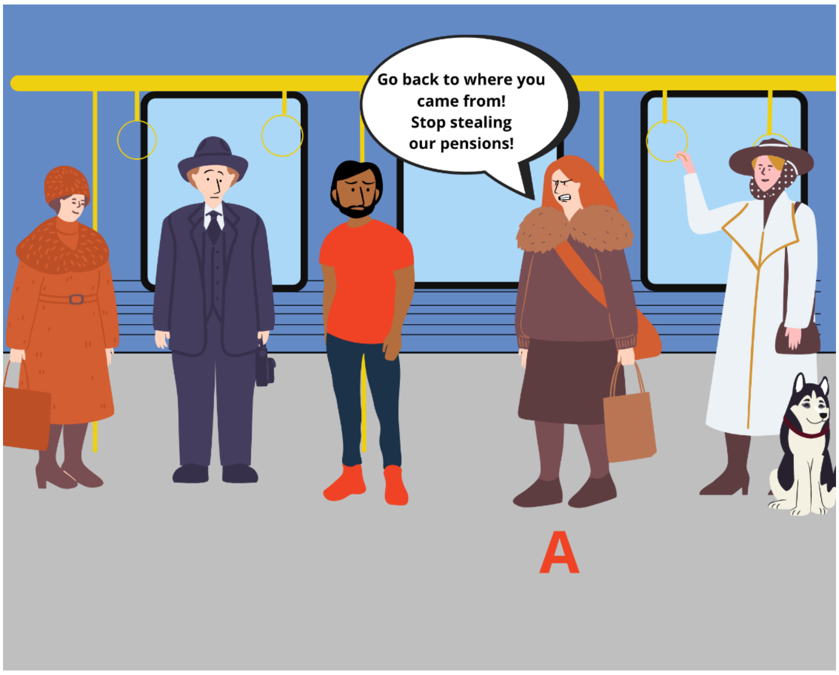
Collective Scenario with 1 silent bystander and 2 opposing (Street)



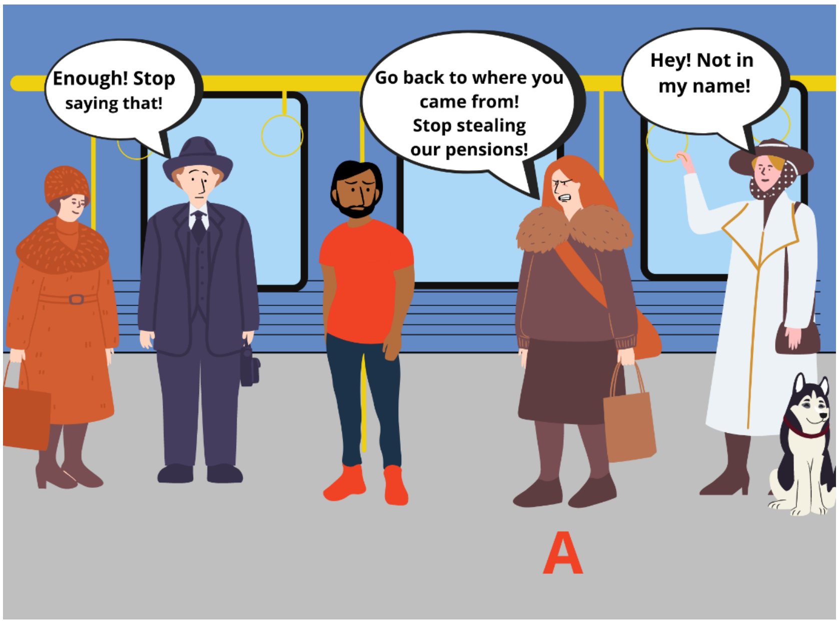
Individual Scenario with 1 silent bystander (Subway)


Go back to where you came from! Stop stealing our pensions!
Enough! Stop saying that!
A

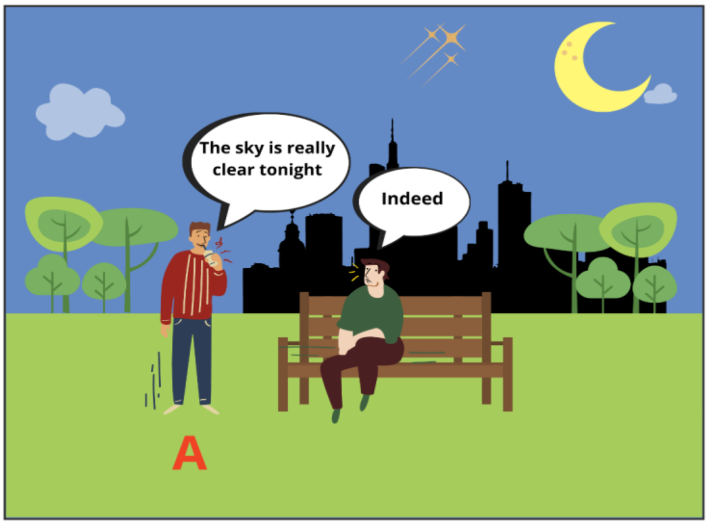
Individual Scenario with 1 opposing bystander (Subway)



Collective Scenario with 3 silent bystanders (Subway)


Enough! Stop saying that!
Go back to where you came from! Stop stealing our pensions!
Hey! Not in my name!
A

Collective Scenario with 1 silent bystander and 2 opposing (Subway)



Visual vignette for attention task

3.3.2. Experiment 2.

3.3.2.1. Study Description.

Based on Experiment 1's findings, we changed to focus exclusively on collective settings, where there were always three bystanders. We varied the number of opposing responses from zero to three (of three total bystanders). This meant that opposition could be absent (0/3), be a minority response (1/3), be a majority response (2/3), or be unanimous (3/3). Additionally, we designated one of the bystanders the "target bystander" so that the questions could focus participants' attention on a specific bystander, and we could thus ask participants to rate the specific target's contribution to overall harm during the hate speech incident. The target bystander could be silent or opposing, with or against the majority.

Accordingly, this yields a factorial design combining two independent variables (IVs). IV1 ("target response") the target-bystander's response to the hate speech incident with two levels: showing opposition vs remaining silent; and IV2 ("majority response") the response of the bystanders majority also with two levels: showing opposition vs remaining silent. Table 2 lists all the conditions. Target bystanders, which are the focus of questions about specific bystanders' contributions to harm, are indicated with an arrow (→).

Table 2. Experimental conditions tested in Experiment 2, with target bystanders indicated with arrows.

Experimental conditions		**Majority response**	
		Remain silent	Show opposition
Target-bystander response	Remain silent	A →🤐🤐🤐	C →🤐😮😮
	Show opposition	B →😮🤐🤐	D →😮😮😮

The factorial design allowed us to test individual (target bystander) and collective (group of bystanders) reactions to hate speech incidents and simultaneously to test the effect of the target bystander responding with or against the majority. As shown in Table 2, the target bystander remained silent in scenarios A and C. However, in scenario A, she did it jointly with all other bystanders, while in C, she remained silent when the majority showed opposition to the hate speech incident. Likewise, in scenarios B and D, the target bystander opposed the attack. Still, in scenario B, she opposed the attack when the majority remained silent. In contrast, in scenario D, she opposed the hate attack together with the rest of the bystanders.

A within-subjects design allowed all participants to evaluate four experimental conditions with zero, one, two or three opposing bystanders (referred to as scenarios A, B, C and D). We collected two dependent variables: the overall level of harm of the incident ("harm" DV1) and the specific contribution of the target bystander to that harm ("contribution" DV2) . For the latter, we asked participants whether the target bystander's response increased or decreased the harm caused by the incident. The order of presentation of these questions was randomised.

We tested the following hypotheses about how bystander responses will affect the perception of the harm caused by a hate speech incident:

H1: *Target-bystander's opposing response* will contribute negatively to (i.e., reduce) the perceived harm. (DV2 as a function of IV1).

H2: When most bystanders remain silent, a *target bystander opposing the attack* will reduce the perceived harm less than when the others oppose the attack (DV2 as a function of IV1 × IV2).

H3: *The level of harm of the incident* will be reduced accordingly to the number of opposing bystanders present. (DV1 as a function of the number of bystanders).

H4: *The level of harm of the incident* will be reduced when showing opposition is the majoritarian reaction among bystanders (DV1 as a function of IV2, indicating a social norm).

H5: *The level of harm of the incident* will be reduced when showing opposition is unanimous among bystanders (DV1 as a function of unanimity, indicating a robust social norm).

3.3.2.2. Participants.

As we planned to analyse the DVs using cumulative link mixed-effects models, we conducted a power calculation through simulation for mixed models with the mixed-power R package (Kumle et al., 2018). In addition, we used pilot data to obtain estimates for fixed and random effects. Results showed that to reach a power of 0.80, 225 participants were required.

We recruited 272 British English-speaking participants through Prolific (www.prolific.co). Prior to analysis, data from four participants who failed the attention checks (see below) were removed. The final sample size included in the study was N=269 participants (134 female, 2 prefer not to say/non-binary).

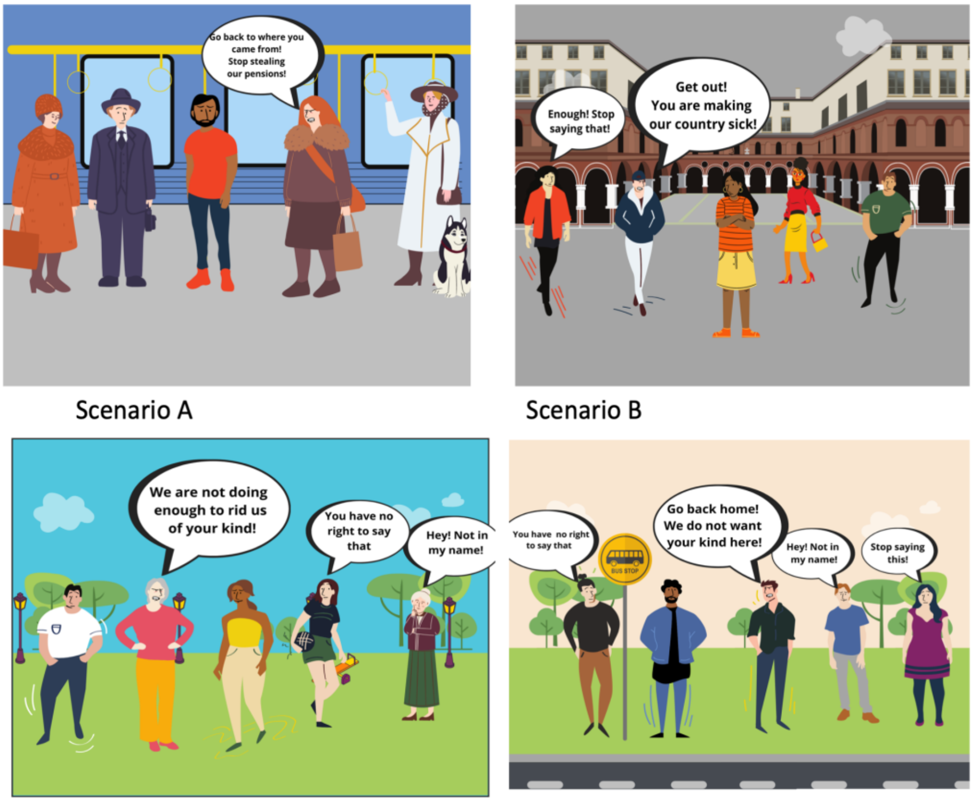
3.3.2.3. Testing Materials.

As for Experiment 1, we created colourful cartoons representing hate speech incidents. However, this time all four scenarios were collective (group) scenarios of three bystanders showing opposition or remaining silent.

Examples of the visual vignettes are shown in Fig. 5. In addition, in each scenario, there was a "target" bystander who either appeared silent or showed opposition, with or against the majority (See Table 2).

Fig. 6 shows, as an example, the target bystander in Scenario D, who appears to show opposition in line with the majority. We showed participants a vignette showing only the target bystander when we asked them to assess a specific target's contribution to overall harm during the hate speech incident (See Supplementary Information section for the full battery of visual vignettes).

Figure 5. The image shows example visual vignettes for each of the 4 experimental conditions: Scenario A with 0 opposers, Scenario B with one, Scenario C with 2 and Scenario D with 3 opposers.


Go back to where you came from! Stop stealing our pensions!
Enough! Stop saying that!
Get out! You are making our country sick!
We are not doing enough to rid us of your kind!
You have no right to say that
Hey! Not in my name!
You have no right to say that
BUS STOP
Go back home! We do not want your kind here!
Hey! Not in my name!
Stop saying this!

Scenario A

Scenario B

Scenario C

Scenario D

Figure 6. The image illustrates the target bystander in Scenario D. Such an image was presented alongside all questions about a bystander's individual contribution to the overall harm caused by the incident to ensure that participants knew which bystander was the focus of each question.

Hey! Not in my name!

3.3.2.4. Attention-check Task.

As in Experiment 1, we used an attention check that appeared randomly in the trial order. It showed 3 characters, two of whom were talking friendly. We asked participants how many people were speaking in the scene, and they had to respond 2 on a 7-point Likert scale to pass the attention check.

3.3.2.5. Procedure.

We conducted the study using the Qualtrics software (www.qualtrics.com). After giving consent, participants were shown four experimental scenarios and one attention check in random order. In each of the experimental trials, a perpetrator shouts a hateful remark towards a victim in the presence of a group of three bystanders who respond individually or collectively, each remaining silent or voicing their opposition against the attack, as we shown in Table 2.

Participants were asked to rate all scenarios regarding the incident's overall level of harm (*"In your opinion, how harmful is the incident shown above?", DV1*). As in Experiment 1, responses were on a 7-point Likert scale ranging from 0 (Not at all) to 6 (Extremely). In addition, we asked them to rate a target-bystander's individual contribution to the harm caused by the incident ("*To what extent does this person's reaction contribute to the harm caused by the incident. If you consider his reaction plays no role, please, place the cursor on zero.", DV2*). To answer this question, we presented participants with a picture of a target bystander (Fig. 6), and they responded using a bipolar 7-point Likert scale ranging from -3 (Reduces the harm) to 3 (Increases the harm). The middle point (0) was explicitly labelled as "neutral" to highlight to participants that this means "had no effect on overall harm".

Participants finished the study by answering basic demographic questions. All participants who completed the survey and did not fail the attention check were paid £0.75 for a maximum of 5 minutes of work.

3.3.2.6. Analysis Strategy.

First, we pre-processed the data by excluding participants who failed the attention check. Then, we ran a series of cumulative link mixed-effects regressions (CLMM, R package "ordinal", Christensen, 2022) to test the hypotheses. All data analyses were performed in R. The OSF repository for this study:

https://osf.io/nfyg9/?view_only=3fe4e0bf7ddd41d4a27dc252cfb67455 contains the data and analyses.

The models reported below (Table 3) were not pre-registered, but the full R analysis script is available in the above study repository. Regression coefficients are reported with 95% confidence intervals (CIs).

Table 3: Cumulative link mixed models tested in Experiment 2.

Model	Outcome variable	Predictor variable
1	Target-bystander's contribution to the harm caused (increases or reduces)	Target-bystander's reaction (show opposition or remain silent)
2	Target-bystander's contribution to the harm caused (increases or reduces)	Target-bystander's reaction (show opposition or remain silent) * social norm (reaction followed by the majority of bystanders)
3	Level of harm of the incident	Number of opposing bystanders (3, 2, 1, 0)
4	Level of harm of the incident	Showing opposition as majority response (social norm supported by the majority)
5	Level of harm of the incident	Showing opposition as unanimous response (robust social norm unanimously supported)

3.3.2.7. Results.

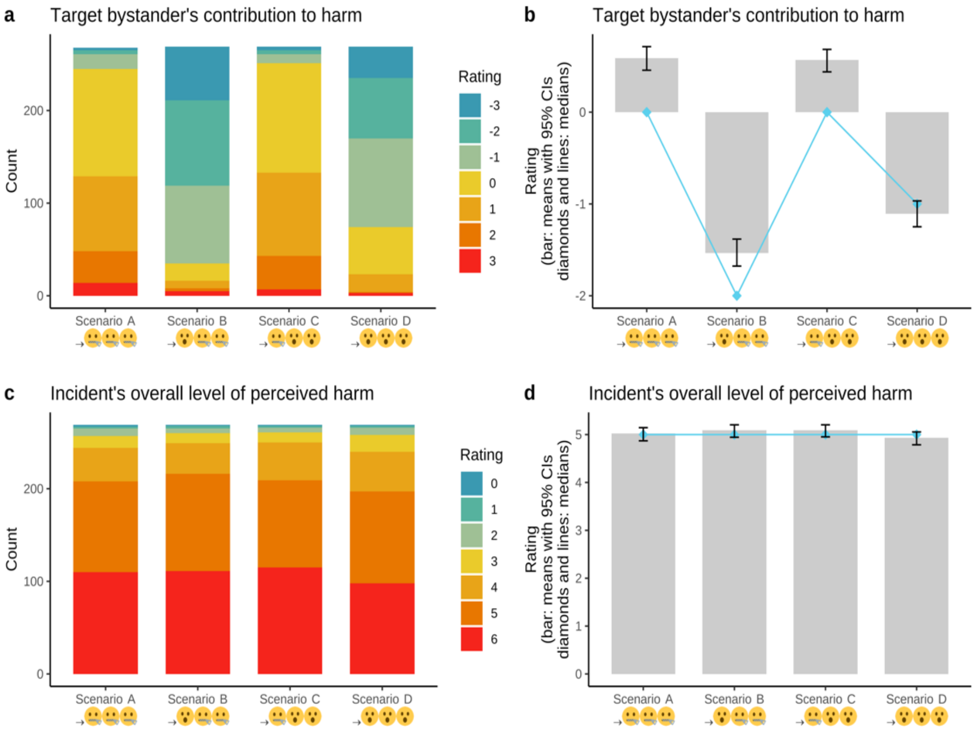
3.3.2.7.1. The specific contribution of the target bystander to that harm.

First, we tested the effect of a target-bystander reaction (showing opposition or remaining silent) on perceived harm. To do so, we ran a cumulative link mixed-effects regression (Model 1) with the target bystander's contribution to the harm caused (reduce or increase) as the outcome variable and the target-bystander reaction (opposing or remaining silent) as the predictor. Results showed that the target-bystander response significantly and negatively predicted harm (i.e., reduced it) when the target bystander opposed the attack (b = -3.57 [-3.90, -3.24], SE = 0.17, t = -21.03, p < .001), confirming H1.

We ran a cumulative link mixed-effects regression (Model 2) with the target bystander's contribution to the harm caused (reduce or increase) as the outcome variable, with this regressed on the target-bystander reaction (remain silent or show opposition), the social norm followed by the majority of bystanders (opposing or remaining silent), and an interaction term. Results showed a nonsignificant effect of the social norm (remaining silent: b = -0.01 [-0.32, 0.29], SE = 0.16, t = -0.07, p = .947); a significant negative effect of showing opposition as target-bystander reaction (reducing harm: b = -3.23 [-3.61, -2.84], SE = 0.19, t = -16.34, p < .001), and a significant interaction between showing opposition as the targeted-bystander reaction and remaining silent as the social norm: (b = -0.83 [-1.26, -0.39], SE = 0.22, t = -3.68, p < .001). Thus, the target bystander's opposition to the attack reduces harm more when it goes against a social norm of being silent, counter to H2, which predicted the opposite effect.

Using the R package "performance", we tested the fit of both previous regression models as indexed by the Bayesian Information Criterion (BIC). The results indicated that Model 2 fits the data better than Model 1 (BIC model 1 = 3236, BIC model 2 = 3223, ΔBIC = 13, weight favouring model 2 = 0.9989). Thus, the best available description of the data is that participants perceived the harm-reducing effect as higher in Scenario B, where a single bystander shows opposition while all the others remain silent (Fig. 7a, b).

Figure 7. Responses are grouped by scenario, and the target bystander in each scenario is indicated with an arrow. (a) Stacked bar chart showing the rating distribution for the target bystander's contribution to harm (positive ratings = increase harm, negative ratings = reduce harm, zero = makes no difference). (b) Grey bars show the mean rating (and whiskers show 95% bootstrapped CIs) of the target bystander's contribution (increase or reduce) to the damage caused by the incident; blue diamonds and lines show median responses. (c) Stacked bar chart showing the rating distribution for the incident's overall level of perceived harm. (d) Grey bars show the mean rating (and whiskers show 95% bootstrapped CIs) of the incident's overall level of perceived harm; blue diamonds and lines show median responses.



3.3.2.7.2. The overall level of harm of the incident.

Secondly, we tested—again, always in collective settings—the effect of several predictors (number of opposing bystanders {0, 1, 2 or 3}, a majority opposition response and a unanimous opposition response) on participants' perceptions of the overall harm caused to victims. For this purpose, we ran three different cumulative link mixed model regressions.

Model 3 regressed the incident's overall perceived harm on the number of opposing bystanders and showed a significant negative effect of the number of opposers (b= -0.16 [-0.157, -0.156], SE < .001, t = -613.42, p < .001). Model 4 had the same outcome variable but regressed this on the majority bystander response (social norm) and showed a nonsignificant effect when the majority opposed (b = -0.24 [-0.53, 0.05], SE = 0.15, t = -1.64, p = .102). Finally, Model 5 regressed the same outcome variable on the dichotomous unanimity variable (whether all bystanders opposed or not, with the former reflecting a robust social norm). The results showed a significant negative effect when all bystanders opposed (b = -0.63 [-0.97, -0.30], SE = 0.17, t = -3.74, p < .001).

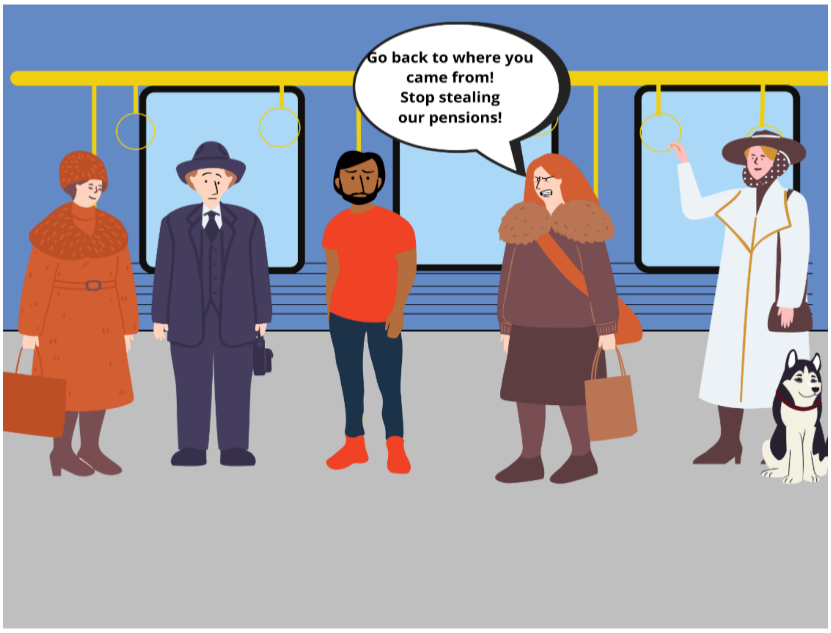
Lastly, using the "performance" package, we evaluated the fit of the three previous regression models (Model 3 BIC = 2077, model 4 BIC = 2071, model 5 BIC = 2067, ΔBIC = 4 for model 5 vs next-best model 4, weight in favour of model 5 = 0.831). Thus, the best available description of the data is that the incident's overall level of harm is better reduced when the opposition against a hate speech incident is unanimous among bystanders, thereby becoming a robust social norm (Fig. 7c, d).

3.3.2.8. Discussion.

Experiment 2 placed a given bystander’s response to a hate speech incident in the context of other bystanders' reactions (reflecting overall levels of opposition/social norms). Results show that participants, as third-party observers, judged that remaining silent could increase the perceived harm of a hate speech incident, that a given individual’s speaking out is more impactful when the majority of bystanders are silent. Crucially, however, the best way to reduce harm overall is to have a robust social norm in favour of speaking out against hate speech. Thus, assessing a bystander's response to hate speech without considering the social context (and any empirical social norms in place) could overestimate its impact on perceived harm. As Fig. 7 shows, the variation in the incident's overall level of harm is relatively small (Fig. 7d) compared to the variation in how a bystander’s response impacts overall harm (Fig. 7a, b) when it is assessed individually. Moreover, although participants praise single opposers who raise their voices amid the silent majority, our results show that only unanimous opposition significantly reduces the public perception of the harm caused.

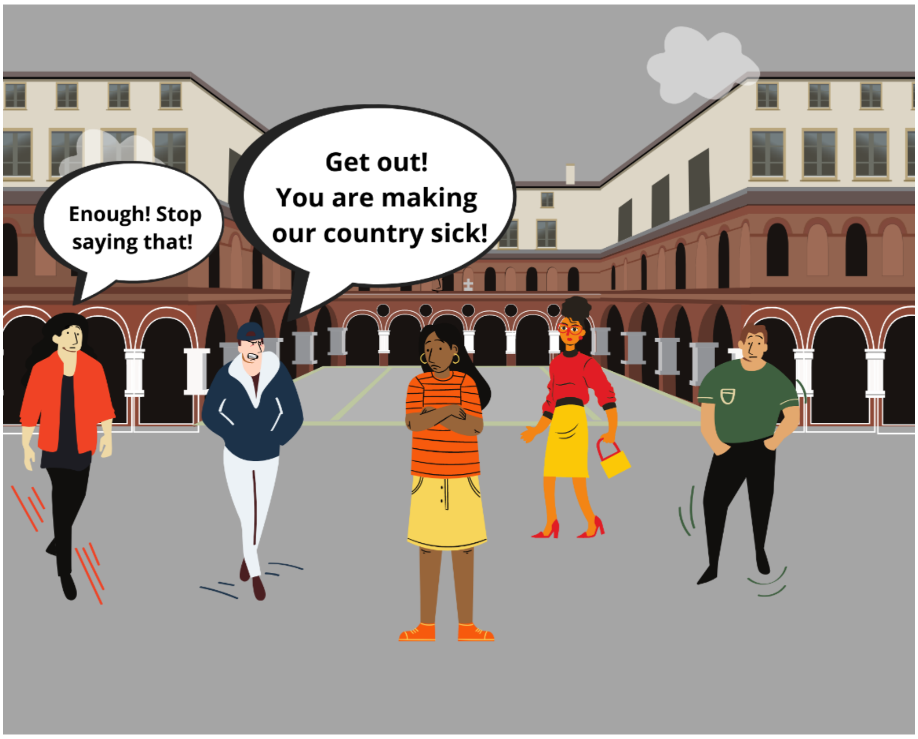
3.3.2.9. Supplementary materials

Testing Materials



Majority response: Remain silent (3/3)

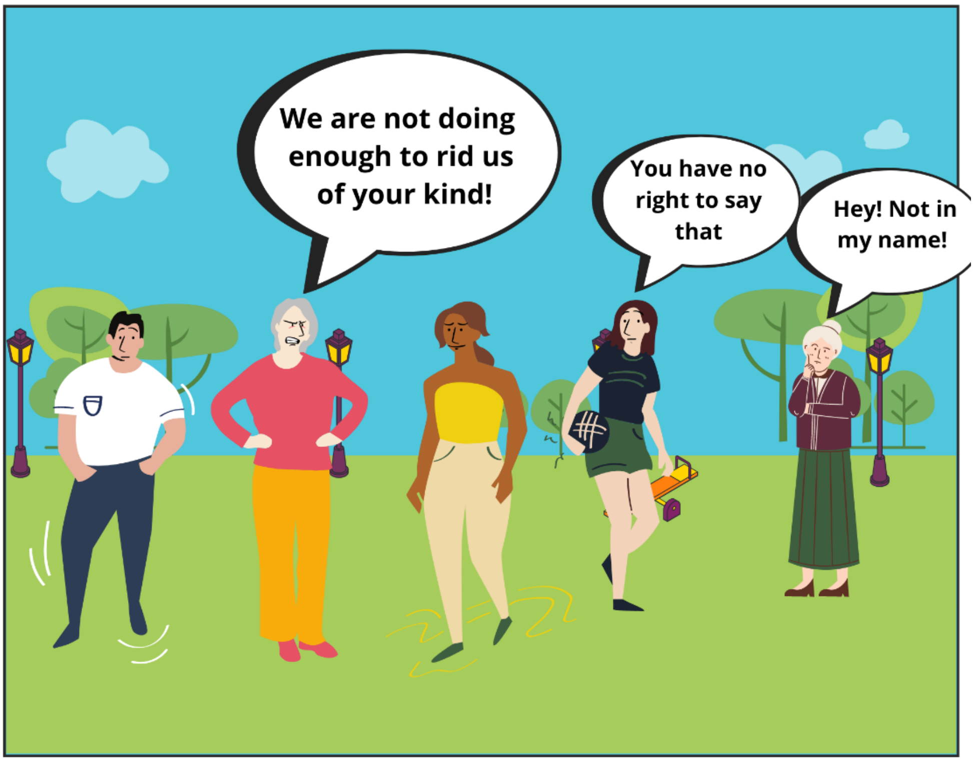
Target-bystander response: Remain silent


Enough! Stop saying that!
Get out! You are making our country sick!

Majority response: Remain silent (2/3)

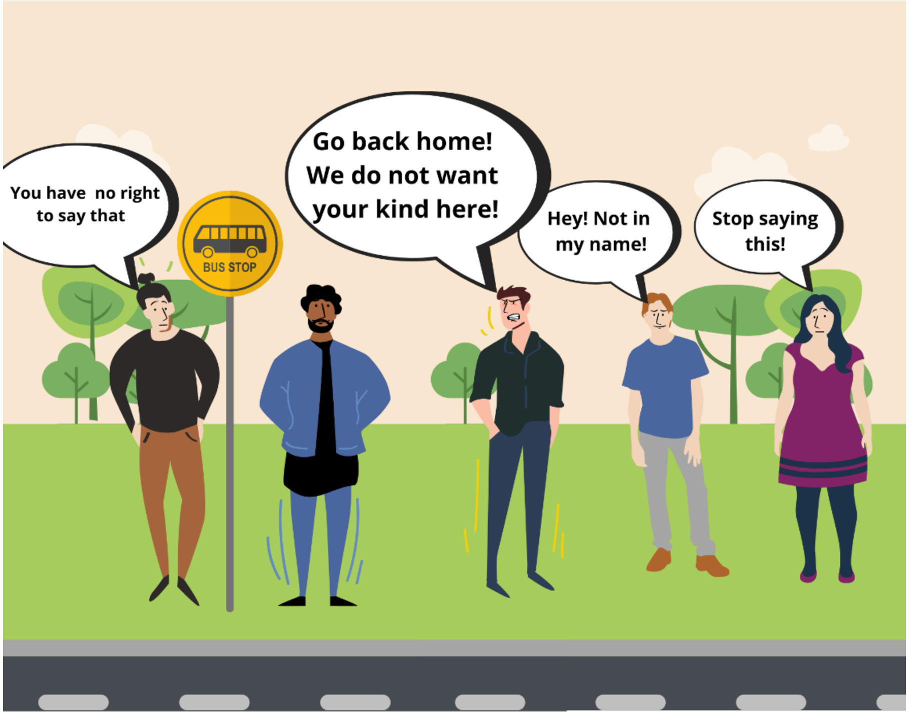
Enough! Stop saying that!

Target-bystander response: Showing opposition


We are not doing enough to rid us of your kind!
You have no right to say that
Hey! Not in my name!

Majority response: Showing opposition (majority 2/3)

Target-bystander response: Remain silent


You have no right to say that
BUS STOP
Go back home! We do not want your kind here!
Hey! Not in my name!
Stop saying this!

Majority response: Showing opposition (3/3)

Hey! Not in my name!

Target-bystander response: Showing opposition

Have you seen the new Marvel film?
Yeah, I liked it

Visual vignette for attention task

3.3.3. General discussion.

Experts from different disciplines have strongly advocated for counterspeech as a tool against hate speech and its harmful consequences for victims and society (for an overview, see Cepollaro et al., 2023). In this paper, our starting point was to explore whether those who might counter or block hate speech find voicing opposition helpful in reducing the harm created.

Our results show that ordinary people overlook the effect of a silent or an opposing response in the harm created by hate speech when they assess those reactions as individual responses from a single bystander. Moreover, opposing a hate attack when all other bystanders keep quiet is seen as more helpful in reducing harm. However, when we offer participants scenes with a social context and a clear social norm against hate speech (followed by most bystanders), they judge that an isolated opposing response does not reduce the perceived harm, though a unanimous collective opposition can do so. Our results support that group responses to hate speech can modulate its damage by indicating either a condoning or a condemning social norm.

Chater & Loewenstein (2022) pointed out that discrimination is a type of social problem, as inequality or misinformation are, in which the phrase "small changes can make a big difference" does not apply. Our results point in the same direction, suggesting that showing opposition against hate speech is ineffective in isolation and that groups need to respond against demeaning and discriminatory speech as a social norm to effectively reduce its harm.

3.3.4. Limitations to Generality.

As hate speech is highly context-dependent, we conducted our study with only British English-speaking participants; further research is needed to explore whether our findings are replicated with non-English-speaking participants from different countries. Likewise, we only tested racist hate speech with case vignettes representing “real-life” attacks. However, future research can extend our findings by investigating people’s responses to hate speech based on different biases (homophobia, transphobia, based on religious hatred, among others) in various settings like online forums.

In addition, our visual stimuli only used counterspeech that confronts perpetrators (e.g., “Stop saying that”, “You have no right to say that”), and further research should explore whether people’s responses change if we direct the counter-speech to the victim (e.g., "Don't believe him", "I welcome you to this country") or modulate it, making it more indirect (e.g., "I am calling the police").

Finally, following our account, in forthcoming work, we will test whether using expressions that imply a collective response would reduce the harm better than those that suggest individual responses (e.g., "We welcome you", "We will report this to the police", "We don't share that opinion").

3.3.5. Data & code availability

The data sets generated and analysed for this study are available through the Open Science Framework, which also contains the analysis scripts and study stimuli: https://osf.io/nfyg9/?view_only=3fe4e0bf7ddd41d4a27dc252cfb67455

3.3.6. Ethical approval.

This study was performed in line with the principles of the Declaration of Helsinki. The Ethics Committee at the LMU approved the protocol for this study (ID-Number 131874 from 10.02.2022).

3.3.7. Informed consent.

All participants provided informed consent before taking part in the study. They received relevant information about the research aim, procedure, duration, and compensation. Furthermore, we informed them that although some visual scenes could be distressful, participating in the experiment would involve no other expected risks and that they could withdraw from it at any time without further consequences.

3.4. Conclusion.

In two experiments, we found evidence that bystanders' reactions when facing a hate speech attack can play a pivotal role in how people view the harm caused. Third-party observers perceived these incidents as causing less harm to victims and society when the implicit social norm (followed by most bystanders) was to show opposition against hate speech, mainly when it was strongly supported (unanimous). We propose that hate speech is better addressed by collective responses than individual efforts. As hate speech attacks are ultimately about demeaning social groups more than specific individuals, they also require collective responses and clear social norms that regulate our coexistence within democratic principles, like tolerance and respect for diversity. Our findings show that people's folk intuitions point in the same direction.

CHAPTER 3. ASSESSING BYSTANDER'S RESPONSES TO HATE SPEECH: COLLECTIVE OPPOSITION IS MORE EFFECTIVE THAN INDIVIDUAL CONFRONTATION

Chapter 4

Revisiting hate speech harm to characterise the speaker.

4.1. Revisiting hate speech harm.

In this section, we revisit our philosophical assumptions regarding hate speech in light of our empirical work.

In Chapter 2, we showed that in those cases where external circumstances prevent hate speech from causing negative consequences to direct targets (e.g., the target was deaf and could not hear a word coming from the speaker), people keep assessing the incidents involving hate speech as highly harmful. We interpret that finding as people recognising hate speech's harm potentiality to perform various speech acts simultaneously (e.g., demeaning targets while encouraging like-minded fellows to act) (Lewiński, 2021), which may harm different people at a time (e.g., direct targets, random bystanders and beyond). Folk intuitions capture this inherent characteristic of hate speech that keeps it harmful when the speakers do not manage to harm their direct targets, something other acts of hate (e.g., bodily actions) cannot achieve.

Then, in Chapter 3, we reported that the harm created by one speaker performing a hate speech act was not affected by the presence of one bystander who voiced opposition. Moreover, when we presented participants with scenarios representing identical hate speech incidents, with either zero, one, two or three opposing bystanders, the harm reduction they perceived was not gradual. What modulates how harmful the hate speech incident was perceived to be was not the severity of the speech itself (e.g., more hostile or violent), which remained the same, but whether bystanders reacted against it or not: a strongly majoritarian response made people perceive the incident as less harmful.

Our findings made us shift our focus from the content (hate speech) to the subject (the speaker) who performs hate speech acts.

Some recent studies exploring the contextual determinants of offensive speech (a variety of harmful speech) argue that the contours of offensive speech are not informed primarily by linguistic cues but rather by a broader contextual assessment that draws retrospectively on information about the speaker's background and identity (Almagro et al., 2022).

Target-group membership primarily drives an utterance's perceived offensiveness (Almagro et al., 2022; Galinsky et al., 2013; Gibson et al., 2020). The same term is perceived as less offensive when the speaker shares group membership with the target and more offensive when the speaker and the target belong to different groups (Almagro et al., 2022; Galinsky et al., 2013; Gibson et al., 2020; Whitson et al., 2017); although participants tend to discount the relevance of those membership norms when asked about the reasons behind their judgments (Almagro et al., 2022).

Similarly, when reclamation[3] motives are inferred in speakers who self-label themselves with slurring terms, people perceive such terms as less insulting or demeaning (Gibson et al., 2020). In those cases, the insulting connotation of a term varies depending on who uses it and for which purposes (Fassoli et al., 2019; Wang et al., 2017).

Then, if the content is not the most powerful in a hate speech act, what is it? Where lies its power to harm?

Let us take a straightforward example. You take a crowded city bus in London to work wearing your favourite light green pullover, and while there, another passenger glares at you and yells: "You have no respect for our national dress code! You are destroying our culture!" You are British and fully aware that there is no such code prescribing not to wear light green clothes, so just move ahead, smiling or amazed. Now imagine that you take the same bus, for the same purpose, wearing a traditional Abaya and while in there, another passenger yells at you: "You have no respect for our country! You are destroying our culture!" You are a Moroccan Muslim living in London. Didn't you find the first case anecdotal, whilst the second resembles a hate speech incident? Why?

Beyond the linguistic content that both similar phrases convey, you can be pretty sure that the first case expresses a personal, individual opinion or belief. In contrast, the second describes a group belief, a widespread and shared belief we can, unfortunately, find in many

[3] Linguistic reclamation or reappropriation is a practice through which members of a marginalised group manage to defeat stigma through self-labelling with slurring terms previously used to target them (Popa-Wyatt & Wyatt, 2018). Reclamation of slurring terms such as the "N" word or "Queer" are paradigmatic examples.

societies. In the second case, you might not be able to identify with precision who else shares the speaker's discourse or how big is the speaker's group. Still, you are confident they exist, maybe even on the same bus. So far, from anecdotal, you find the second incident stressful, demeaning, violent, and harmful.

In the second case, we support the idea of the speaker performing a hate speech act with a plural self-awareness, knowledge and commitment to being many in doing so (Schmid, 2014a, 2014b). Correspondingly, the audience receives that xenophobic message as a group speech in a way that is different from how this audience gets the message against wearing a light-green pullover, and that could explain, at least partially, why hate speech acts seem to be hardly blocked or countered individually, as we reported in Chapter 3.

A hate speech act communicates to its audience a group belief resulting from the complex interaction between the biases captured from external social cues and the subject's decision to adopt them (as its own), reinforcing those biases and prejudices as a social norm (Ayala-López, 2018). In addition, passively tolerating or actively supporting the performance of hate speech acts may normalise them (Ayala & Vasilyeva, 2016). Once that happens, the plural subject of those actions is guaranteed, bigger or smaller, depending on whether the social norm is strongly followed, but certainly plural.

We defend the idea that performing actions like demeaning, subordinating and ranking groups of people as inferior (Langton, 2018a) is only effective if people perceive that there is a group who shares and supports the performance of such actions. In other words, hate speech's capacity to harm people lies, at least partially, in the assumption that the beliefs and attitudes expressed by such a speech are shared by more than one individual.

When participants in our study (Chapter 3) assessed bystanders' opposition against a hate speech act and reported a group opposition was needed to counter the harm created effectively, they intuitively perceived hate speech as being uttered in a *we-mode*. This interpretation of our findings makes us venture a new characterisation for speakers in hate speech acts, which challenges the idea that they are merely isolated individuals or "lone wolves" and instead casts them as group members.

4.2. A lone wolf or a group member?

As we mentioned in Chapter 1, speakers alone do not have absolute control over the speech acts they perform. Among other factors, they do not control who makes up their audience or how this audience perceives or influences the resultant (hate) speech act.

Does the audience of hate speech perceive it as coming from an isolated individual or a group member?

Whether the audience of hate speech perceives it as coming from an isolated individual or a group member is, to our knowledge, unexplored in the hate speech literature. Despite being vast, the instructive and insightful research around hate speech acts in the last two decades has been primarily theoretical, with limited empirically informed proposals (See Almagro et al., 2022; Cepollaro, 2023, and Zapata & Deroy, 2023 for an overview). Moreover, most of the examples analysed have involved one isolated speaker who intends meaning and an audience of one or more listeners who attempt understanding, which could have masked a fact: the speaker may not be an isolated individual but a group member (Hughes, 1984).

In order to shed some light on this feature of hate speech, henceforward, we review the empirical findings on bullying literature. Similarly to hate attacks, bullying attacks can be performed verbally and nonverbally, and the aggressors manage to harm their targets because they are not alone: they count on supporters and accomplices (Harvey et al., 2007; Namie & Lutgen-Sandvik, 2010; Ng et al., 2022). The difference between verbal bullying and hate speech lies in the target. As we mentioned before, we defend the idea that the target in hate speech is a group, a disfavoured collective or a minority, targeted by the speaker based on its identity marks (e.g., being dark-skinned or homosexual); in bullying, nevertheless, the target is an individual, targeted based on his personal characteristics (e.g., being shy or introvert).

Empirical psychological research has extensively studied how bullying victims and bystanders perceive verbal and nonverbal aggressions in the workplace. In a study conducted with a sample of 7740 adults, closely reflecting the U.S. census data, participants reported that multiple aggressors performed a third of bullying cases. In contrast, in nearly 70% of cases, bullies were solo actors, which seemed to evidence that bullies were "lone wolves". However, looking deeper into participants' responses, researchers found that in nearly 60% of the solo-bully cases, respondents believed the bullies received support from upper managers, bullies' peers, and even targets' peers, which supposed that nearly three-quarters of bullying cases were perceived as concerted and collective to some degree (Namie & Lutgen-Sandvik, 2010).

In this and other similar studies, researchers concluded that bullying at the workplace (including verbal harassment) only occurs if bullies feel they have the blessing, support, or at least the implicit permission of superiors and other co-workers to behave in this manner, implying that these acts are mainly perceived as collective (Harvey et al., 2007; Namie & Lutgen-Sandvik, 2010; Ng et al., 2022).

These studies support the idea that failing to recognise those aggressions as collective worsens its progression and impedes the involved parties (i.e., co-workers, managers and institutions) to respond accordingly, increasing the harm caused not only to victims and bystanders but to the entire workplace (Namie & Lutgen-Sandvik, 2010; Ng et al., 2022). The similarities between verbal bullying and hate speech make us wonder whether we also fail to recognise the collective dimension of hate speech.

In addition, throughout history, groups in conflict have found it advantageous to attempt to marginalise their opponents verbally with disparaging names and slurs (Gibson et al., 2020). Researchers have shown that the recent expansion of the acts of hate (including hate crime, hate group activity and hate speech) is mediated by attitudes associated with an intergroup dispute for social dominance (Hoover et al., 2021; Duckitt & Sibley, 2017; Charles-Toussaint & Crowson, 2010). Moreover, they confirmed that ordinary people see those acts of hate as responses to perceived realistic or symbolic outgroup threats against moral values like loyalty, authority, or purity (Hoover et al., 2021).

By using hate speech, speakers assign their targets a low-power role, altering social norms, perpetuating degrading practices against the target group and reinforcing social hierarchies beneficial to the speaker's group (Ayala-López, 2018; Popa-Wyatt & Wyatt, 2018; Popa-Wyatt, 2021).

Without empirical studies that explicitly explore whether the audience perceives a hate speech act as coming from a solo aggressor or a group member, we argue that the studies reviewed here provide, at least partially, a foundation to characterise hate speech as a mechanism used in inter-group disputes for dominance, either economic, social or moral, supporting the assumption that hate speakers act necessarily as group members.

4.3. Group speakers and the right to freedom of speech.

Which kind of speech deserves to be considered hate speech and combated accordingly is under contention in the literature (Fraser, 2023; Lepoutre, 2021). Within this debate, some distinguish between "inferiorising" and "expulsory" hate speech (Fraser, 2023). The first presents its targets as less than full persons (e.g., "Jews are poison", "Chinese are a virus"); the second attempts to drive its targets out of a political community (e.g., "Go back home!" "You are making our country sick!").

We contend that, in both cases, the targeted group is presented as not having the same value as the speaker's group, as unworthy of social respect. Moreover, in both cases, the message conveyed is "Your group does not belong to mine, and we do not want your group in ours". Such speech presupposes the existence of two groups: the speaker's group and the target's group. Tolerating hate speech erodes the essence of diverse, contemporary societies because its message implicitly fosters precisely the opposite: uniform societies that do not welcome diverse groups of people, or at least do not recognise their rights and citizenship in equal terms.

Economic globalisation has made contemporary societies diverse in many aspects due to the increasing interdependence among nations and cultures. Citizens might not share a common personal or cultural background in such societies. Therefore, a strong battery of individual rights and freedoms ensuring egalitarian treatment and a tolerant and inclusive public dialogue becomes essential (Lepoutre, 2021). In that context, Freedom of Expression is undoubtedly a solid defensive wall against censorship and dogmatism for diverse societies.

However, we should remember that Freedom of Expression was primarily conceived to protect individuals' right to share ideas and thoughts freely. It was granted with superior protection to ensure that each individual, as a citizen, can take part in the social dialogue in equal conditions (Dworkin, 1977; Scalon, 1972), not to protect the promotion of group speech that targets minorities or disfavoured collectives advocating precisely the opposite: segregation, apartheid and discrimination.

As mentioned above, hate speech acts exclude people, rank them as inferior (Langton, 2018a), based on their identity marks, and do that with normative aspirations. They create unfair hierarchies and threaten a peaceful coexistence (Benesch, 2013; Delgado, 1993; Lawrence, 1993; Matsuda et al., 1993). Therefore, they look closer to cases of undermining propaganda, as they promote unjust world orderings in our conversational common ground (cf. Stanley, 2015), rather than an individual expression of rejection or fear. Freedom of speech principle will hardly defend our societies of such discourses.

Are we suggesting that freedom of speech cannot be exercised jointly with like-minded fellows as in a demonstration defending human rights? Not at all. In a demonstration, a group of people exercise two rights, their right to free speech and free association, which are distinct and protected at different levels. Simplifying greatly, in the context of a demonstration, one protects *what* is said, the other one protects *who* says it.

Freedom of speech is granted with special protection for the reasons mentioned above. In contrast, freedom of association has more limitations based on, for example, public order concerns (to name one of the most important). Those limits on freedom of association exist precisely because a single individual sharing a thought is quite different from many individuals doing so. Groups are clearly more powerful than isolated individuals; therefore, freedom of association right has always been the subject of watchful scrutiny. The discussion about this end largely oversteps the scope of the present dissertation. Notwithstanding, based on that distinction, we propose an additional focus of attention when exploring hate speech: Who is the speaker.

Hate speech is frequently presented as the expression of just one individual, which seems difficult to regulate without raising suspicions about a potential conflict with the individual right to free speech. Mainly when hate speech degrades a target but does not explicitly incite violence against it. However, in addition to the doubts regarding its content, we defend the idea that we should worry about who is behind such hateful discourses, and, in case we find the speaker is actually a group, democratically evaluate whether groups (that sometimes include corporations) should be granted freedom of speech in equal terms with individuals.

Recently, such a debate took place in the United States and reached the Supreme Court. In Citizens United v. Federal Election Commission, 558 U.S. 310 (2010), it was discussed whether a non-profit association like conservative Citizens United could broadcast a critical film against a democrat candidate (Hillary Clinton) within the time restriction for doing so (30 days of a primary election). A majority of justices established Citizens United's corporate right to freedom of expression should prevail against that restriction in the same terms as ordinary citizens' civil rights. However, a dissenting opinion by Justice John Paul Stevens (the second-oldest justice in the history of the U.S. Supreme Court and the third-longest serving justice) defended that establishing such a right contravenes the constitutional First Amendment (Stevens, 2010).

Stevens defended that corporations should not be given speech protections under the First Amendment, which protects "individual self-expression, self-realisation and the communication of ideas". He argued that corporations "unfairly influence electoral processes with vast sums of money that few individuals can match, giving the impression of widespread acclaim regardless of actual support of political parties and agendas". Accordingly, he defends that companies' intervention in such political processes needs regulation (Stevens, 2010).

We should be cautious when a group performs hate speech for similar reasons: powerful members of a group supporting hate speech could help those harmful discourses appear as they were a widespread acclaim regardless of actual support. In this context, highlighting the group nature of hate speech is of special relevance: Should harmful group speech with policy aspirations against disfavoured and minoritarian groups be granted freedom of speech protection under equal conditions alongside individual speech?

As we mentioned before, partaking in the debate concerning which group speech should be protected, and which doesn't, largely exceeds the limits of the present dissertation. However, regardless of whether the group is the sum of biased individuals, or includes economic lobbies fighting for power through social confrontation, it must be subject to careful scrutiny. More importantly, besides watching the limits of its linguistic content and the actions performed by hate speech acts, we should watch whether the principle of freedom of expression needs to be restricted based on who performs such harmful speech.

4.4. A collective response to a group speech.

Consistent increment of hate incidents on both sides of the Atlantic (Eligon, 2018; Levin & Reitzel, 2018; Myers & Lantz, 2020) might make us wonder whether we efficiently respond to hate speech acts. The policies implemented so far are focused on punishing individual speakers with criminal or civil sanctions and encouraging bystanders to take action individually by showing opposition or denouncing the incidents to competent authorities (in countries with hate speech regulations). However, addressing individual behaviours rather than the system in which those individuals operate has been pointed as the root of behavioural public policy astray in phenomena in which the sum of individual efforts seems not to suffice (e.g., climate change, obesity, or pollution from public waste), requiring systemic changes that involve collective actions (Chater & Loewenstein, 2022).

For example, directing citizens to contribute to private pension systems individually seems not to suffice in countries with high unemployment levels and salaries that do not cover life costs. Developing a mandatory public pension system by sustainable taxation seems more suitable there. Similarly, directing individuals to follow a diet against obesity seems inadequate in countries where healthy food is a luxury, not affordable by the majority, and market prices of the most basic food basket can grow without limits (Chater & Loewenstein, 2022).

Would hate speech be a phenomenon of such kind? Are our findings (showing that people's perception of individual efforts against hate speech are unable to reduce the harm it creates) pointing in that direction? Do we need policies against hate speech to promote a duty to respond against it? In which terms?

As previously explained, speech acts are inherently social and interactive, meaning their audience can influence which actions a speaker can perform and to what extent (Kukla, 2023; Sbisà, 2001). According to empirical findings concerning the two phenomena explored here (bullying and hate speech), doing nothing as a bystander is not a neutral act. It lets those harmful actions build their way to a social norm and eventually destroy social coexistence (Ayala and Vasilyeva, 2016; Langton, 2018b; Lepoutre, 2021; Namie & Lutgen-Sandvik, 2010).

Do passive, silent bystanders worsen or buffer the harm created by hate speech acts? To our knowledge, this extent is still unexplored in the hate speech literature (in real-life settings) but has already been explored in bullying studies, which show silent bystanders act as further demands for targets whilst active opposing bystanders act as resources (Ng et al., 2022). Victims of bullying (including verbal harassment) reported that such incidents caused the same hurting feelings of shame, disrespect and social exclusion, regardless of bystanders' responses. Still, when the bystanders present faced the attack and remained quiet, they added a mental overload to the victims by creating ambiguity, making them doubt whether their suffering came as a result of the aggression or a personal extreme sensitivity, whether they misunderstood the speaker words, or whether the speech was legitim and shared by their colleagues present (Ng et al., 2022). According to these findings, targets' mental overload might be higher when they suffer an attack in front of silent bystanders.

This is in line with our findings in Chapter 3. We found that third-party observers of a hate speech incident evaluated the speaker as similarly blameworthy regardless of whether the incident happened in front of silent or opposing bystanders. As mentioned above, third-party observers blamed silent bystanders not as members of the speaker's group but for creating further uncertainty regarding the audience's stance on the attack: whether the audience acquiesced it, supported the hater, disregarded the targets or some combination of those options (Lutgen–Sandvik & McDermott, 2008).

Then, if a silent response worsens hate incidents, should we always voice opposition?

Promoting a collective duty to respond to hate speech incidents can be achieved by finding alternative methods beyond verbal confrontation, which can be counterproductive and inefficient in most daily life incidents. Signalling that we do not team up with the speaker in any available way becomes crucial. Moreover, showing it in a strongly majoritarian way reveals essential too.

Public policies against hate speech should point in that direction, supporting a civil duty to oppose social practices that erode coexistence in diverse societies at a systemic level, pathing the way to the emergence of social norms. Looking around, we find many examples of such norms: not interfering in the way of an ambulance, respecting pedestrian crossings, queueing up, or not smoking in close spaces are some of them.

However, should we respond collectively if a unanimous response seems more efficient in reducing hate speech harm than an individual one? Why? If our conclusions in Chapter 3 point in the right direction, people take random collections of individuals as representative group samples, inferring from bystanders' reactions the social norm in place and how strongly that norm is endorsed. In addition, when people infer that the social norm of responding against hate speech is stronger (i.e., observed in a majoritarian way), people also infer that the group who follows that norm is bigger than the speaker's group, reducing the harmful potentiality of such an act. Therefore, hate speech acts appear to third-party observers as less harmful when they occur in collective settings where the unanimous response is to react against it.

Studies conducted with nationally representative samples in the United States show that witnesses, more often than targets, reported that no one supported bullies in their harming actions. They showed that when targets suffer an attack, and their co-workers seem to look on silently, targets receive that failure to respond as complicity (Namie & Lutgen-Sandvik, 2010). In contrast, bystanders were less likely to perceive others' silence as a type of support (Dillon & Bushman, 2015), which may be linked with primarily unintentional support (Ayala & Vasilyeva, 2016; Langton, 2018b). However, by remaining silent, bystanders inadvertently become speakers' accomplices.

Not allowing ourselves to be counted as part of those who accept or tolerate hate speech sends a message to targets and other bystanders that the speaker's group is not hegemonic. Importantly, according to our findings in Chapter 3, this is a collective task. Individual efforts opposing hate speech have no effect –or minimal effect– in reducing the harm created by such acts. Teaming up to collectively opposing phenomena threatening peaceful coexistence is necessary in diverse societies.

Our studies on collective responses, such as social distancing and vaccination, during the recent COVID-19 pandemic have revealed a key insight: By highlighting the compliance of others in our communities, we can achieve widespread adherence to public policies. This underscores the significant role of social expectations in norm change, a phenomenon that has been demonstrated in numerous laboratory and field experiments (Bicchieri, 2016; Borgonovi & Andrieu, 2020, as cited in Tunçgenç et al., 2021). Therefore, emphasising collectivistic values and the efficacy of collective actions in situations requiring a collective behavioural response showed the most relevance (Tunçgenç et al., 2021).

Drawing on the notion of affordances (opportunities for taking action), researchers have proposed that we, as humans, perceive the opportunities we have to take action only concerning the patterns of behaviour that characterise the form of life in which we are immersed, represented as implicit and explicit social norms (Ayala, 2016; Heras-Escribano, 2019). But importantly, responding collectively to threats that target us as group members and not only as individuals passes through recognising that opportunities for taking action are, in those cases, a collective task. Through "plural self-awareness", we, as a group, can discern the situations or events that put us in a potential "group mode", allowing us to take action accordingly (Schmid, 2023).

Furthermore, when more than one social group shares the social space, our actions (or inactions) can make us be seen as part of a group we do not want to be part of. In such context, a plural self-awareness implies (1) the collective ability to recognise our dual status as individuals and group members; (2) the collective opportunities that we, as a group, have for taking action; (3) our capacity to accept or reject to be seen as part of a group; and (4) the responsibility for failing to take action together, collectively.

Making our status as non-members of the speaker's group explicit and doing it collectively to avoid harming others unintentionally becomes a duty for citizens in diverse societies. We should be responsible for failing to act collectively, independently of whether we are formally constituted groups or just random collections of individuals, like those who share a bus in which a hate speech act is taking place (May, 1990; Petersson, 2008; Schmid, 2018a, 2018b, 2023; Tännsjö, 2007). Our collective responsibility for our actions and omissions as group members comprises our responsibility for failing to get our plural act together in situations or events where, as in hate speech acts, only a collective response can reduce the harm created (Schmid, 2023).

Emphasising the identification of (hate) speakers as group members and recognising a silent response as indirect support to such a group is of utmost importance. Public policies should be oriented towards this, enabling individuals to grasp the extent of the threat posed by hate speech to democracies and the potential strength of a collective and active response against it.

Finally, failing to recognise hate speech as a group speech may carry adverse effects:

1. It could prevent us from responding timely. Assuming the speaker is only a biased individual may add uncertainty on whether such speech is protected by the principle of freedom of expression, one of the foundations of democratic societies. In daily life circumstances, this time-consuming mental operation can prevent us from denouncing perpetrators or supporting victims timely. Moreover, questioning the limits of individual freedom of expression opens the door to censorship advocacy.

2. It could hinder our ability to respond effectively, assuming that responding to such discourses confines us to a one-on-one debate. Research has demonstrated that bystanders of hate incidents are more likely to intervene when they realise the number of offenders and the severity of the consequences (Fischer & Greitemeyer, 2013; Kazerooni et

al., 2018; Taylor et al., 2019). Neglecting the group aspect of hate speech may lead to inaction, which can occur not only in quotidian minor incidents we witness silently but also in election processes. By abstaining from voting, we may miss the opportunity to block political options that tolerate and support discriminatory and subordinating behaviours as social norms, with dramatic consequences for democratic societies.

Along the same line, conceiving an opposing response to hate speech as an individual task may overburden most people. When people (as victims or bystanders) face a hate speech act and perceive it as coming from a group that they are not able to size, they may doubt whether a single opposer, far from countering the hate action, would indeed worsen the situation (i.e., making the harmful bias more salient or putting themselves at risk), which can lead to inaction.

When people oppose hate speech together, they are perceived as group members, regardless of their personal motivations (e.g., whether they identify as part of the target group, are human rights advocates, or believe all extreme speech should be banned). A unified response sends a strong message to targeted victims that hate speech is not socially endorsed. It also communicates to victims and bystanders that tolerance to diversity is the prevailing social norm, with hate speech being the exception.

Chapter 5
Conclusions

We started the present dissertation by arguing that we can harm others with our words. This affirmation seems obvious, but it is not. More often than we imagine, we forget that through our words, we promise, urge and demand but also threaten, punish and harm others. As poet Mario Benedetti said, our words kiss and bite. Therefore, drawing on Speech Act theory, we began our research characterising hate speech as actions performed and defended the idea that most hate speech not only causes harm but constitutes harm to its targets (Langton, 2018a).

In addition, as hate speech demeans and ranks people as inferior (Langton, 2018a) based on identity marks such as ethnic origin, gender, and religion, among others, we alleged that hate speech' targets are groups and not isolated individuals.

Furthermore, we advocated the notion that the audience of hate speech encompasses not only the direct targets but also random bystanders who inadvertently witness hate speech incidents. This led us to the conclusion that anyone, regardless of being a direct target or not, can find themselves part of the audience of a hate speech act, thereby potentially experiencing its harm (Anderson & Barnes, 2022). However, as members of its audience, each of us has the power to respond to the actions it seeks to perform, thereby mitigating the harm it inflicts (Kukla, 2023; Sbisà, 2001).

In Chapter 2, following an empirical moral philosophy approach, we conducted a study which shows that people, as third-party observers, are more averse towards hate speech than physical hate actions when intention and consequences remain equal. Our results extend and complement a range of findings showing third-party' moral evaluation and punishment heuristics are determined not only by intentions and outcomes but also by associations with intrinsic actions' properties and people's aversion to them (Miller et al., 2014).

We confirmed that no lower moral condemnation of verbal hate attacks is ingrained in ordinary people's moral dispositions. Our results show that ordinary citizens are open to recognising the harm caused by words, which is a solid basis for developing public policies that reinforce civic engagement against hate speech incidents.

Moreover, our findings contribute to the discussion regarding the limits of the freedom of speech principle that set them in the actual harm caused to victims, reporting that ordinary citizens clearly recognise hate speech harm lies beyond the negative consequences created for direct targets.

Continuing our empirical research, in Chapter 3, across two experiments, we found evidence that bystanders' reactions when facing a hate speech attack can play a pivotal role in how people view the harm caused.

Experiment 1 shows that bystanders' reactions affect the perception of the harm created by a hate speech attack only in collective settings when other bystanders are shown. This finding may suggest that when we do not offer participants enough elements to intuit the social norm against hate speech (by showing them how other bystanders react), they evaluate both hate incidents as similarly harmful, independently of whether the bystander present responded by remaining silent or showing opposition. Additionally, our results suggest that people evaluate scenarios where a group of bystanders voice their opposition as less harmful than those where a group remains silent.

Experiment 2 placed a given bystander's response to a hate speech incident in the context of other bystanders' reactions (reflecting overall levels of opposition/social norms). Results show that participants, as third-party observers, judge that remaining silent could increase the perceived harm of a hate speech incident and that a given individual's speaking out is more impactful when most bystanders are silent. Crucially, however, the best way to reduce harm overall is to have a robust social norm in favour of speaking out against hate speech. Thus, assessing a bystander's response to hate speech without considering the social context (and any empirical social norms in place) could overestimate its impact on perceived harm. The variation in the incident's overall level of harm is relatively small compared to the variation in how a bystander's response impacts overall harm when it is assessed individually. Moreover, although participants praise single opposers who raise their voices amid the silent majority, our results show that only unanimous opposition significantly reduces the public perception of the harm caused.

We concluded that collective responses better address hate speech than individual efforts. As hate speech attacks ultimately demean social groups more than specific individuals, they also require group responses and clear social norms that regulate coexistence within democratic principles, like tolerance and respect for diversity. Our findings show that people's folk intuitions point in the same direction.

In Chapter 4, based on our empirical findings, we ventured a new characterisation for speakers in hate speech acts. This characterisation challenges the idea that they are merely isolated individuals or "lone wolves" and casts them as group members instead. We defended the idea that performing actions like demeaning, subordinating and ranking groups of people as inferior (Langton, 2018a) is only possible if people perceive that there is a group that shares and supports the performance of such actions.

Accordingly, we characterised hate speech as a mechanism used in inter-group disputes for economic, social or moral dominance, supporting the assumption that hate speakers act necessarily as group members (Hoover et al., 2021). We defended the idea that by using hate speech, speakers assign their targets a low-power role, altering social norms, perpetuating degrading practices against the target group and reinforcing social hierarchies beneficial to the speaker's group (Ayala-López, 2018; Popa-Wyatt & Wyatt, 2018; Popa-Wyatt, 2021).

Moreover, we argued that hate speech is frequently presented as the expression of just one individual, which makes it challenging to regulate without raising suspicions about a potential conflict with the individual right to free speech, mainly when hate speech degrades a target but does not explicitly incite violence against it. However, we defended the idea that we should worry about who is behind such hateful discourses and, in case we find a group carefully scrutinising whether freedom of speech should protect such discourse. Besides watching the limits of its linguistic content and the actions performed by hate speech acts, we should watch whether the principle of freedom of expression needs to be restricted based on who performs such harmful speech.

We finish our dissertation emphasising the collective nature of an effective bystander response. We stress the importance of not allowing ourselves to be counted as one of those who accept or tolerate hate speech. This refusal tells targets and bystanders that the speaker's group is not hegemonic. Responding to hate speech should be encouraged as a collective task since individual efforts opposing hate speech have no effect –or minimal effect– in reducing the harm created by such acts. Therefore, we defended the idea that teaming up to collectively opposing phenomena that threaten peaceful coexistence is necessary in diverse societies. Moreover, showing opposition in a strongly majoritarian way reveals essential.

Public policies aimed at combating hate speech should not only discourage its use but also actively promote a civil duty to oppose social practices that undermine coexistence in diverse societies. These policies can pave the way for the establishment of new social norms, instilling hope for a more inclusive and tolerant future.

References

Agatston, P. W., Kowalski, R. & Limber, S. (2007). Students' Perspectives on Cyber Bullying. *Journal of Adolescent Health*, *41*(6), S59–S60. https://doi.org/10.1016/j.jadohealth.2007.09.003

Alexander, L. A. & Horton, P. (1983). The Impossibility of a Free Speech Principle. *Northwestern University Law Review*, 78(5), 1319–1357.

Alfano, M., Machery, E., Plakias, A. & Loeb, D. (2022). Experimental Moral Philosophy. In E. N. Zalta & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/experimental-moral/

Allen, M. (2017). *The SAGE Encyclopedia of Communication Research Methods*. SAGE Publications, Inc. https://doi.org/10.4135/9781483381411

Almagro, M., Hannikainen, I. R. & Villanueva, N. (2022). Whose Words Hurt? Contextual Determinants of Offensive Speech. *Personality and Social Psychology Bulletin*, 48(6), 937–953. https://doi.org/10.1177/01461672211026128

Álvarez-Benjumea, A. (2023). Uncovering hidden opinions: social norms and the expression of xenophobic attitudes. *European*

Sociological Review, 39(3), 449–463. https://doi.org/10.1093/esr/jcac056

Anderson, L. & Barnes, M. R. (2022). Hate Speech. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Spring 2022). https://plato.stanford.edu/entries/hate-speech/

Ashburn-Nardo, L., Blanchar, J. C., Petersson, J., Morris, K. A. & Goodwin, S. A. (2014). Do You Say Something When It's Your Boss? The Role of Perpetrator Power in Prejudice Confrontation. *Journal of Social Issues*, 70(4), 615–636. https://doi.org/10.1111/josi.12082

Ashburn-Nardo, L., Lindsey, A., Morris, K. A. & Goodwin, S. A. (2020). Who Is Responsible for Confronting Prejudice? The Role of Perceived and Conferred Authority. *Journal of Business and Psychology*, 35(6), 799–811. https://doi.org/10.1007/s10869-019-09651-w

Atzmüller, C. & Steiner, P. M. (2010). Experimental Vignette Studies in Survey Research. *Methodology*, 6(3), 128–138. https://doi.org/10.1027/1614-2241/a000014

Austin, J. L. (1962). *How to do Things with Words*. Oxford University Press.

Ayala, S. (2016). Speech affordances: A structural take on how much we can do with our words. *European Journal of Philosophy*, *24*(4), 879–891. https://doi.org/10.1111/ejop.12186

Ayala, S. & Vasilyeva, N. (2016). Responsibility for Silence. *Journal of Social Philosophy*, *47*(3), 256–272. https://doi.org/10.1111/josp.12151

Ayala-López, S. (2018). A Structural Explanation of Injustice in Conversations: It's about Norms. *Pacific Philosophical Quarterly*, 99(4), 726–748. https://doi.org/10.1111/papq.12244

Bach, K. & Harnish, R. (1979). *Linguistic Communication and Speech Acts*. MIT Press.

Barendt, E. (2009). Incitement to, and Glorification of, Terrorism. In I. Hare & J. Weinstein (Eds.), *Extreme Speech and Democracy* (pp. 445–462). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199548781.003.0023

Barhight, L. R., Hubbard, J. A., Grassetti, S. N. & Morrow, M. T. (2017). Relations Between Actual Group Norms, Perceived Peer Behavior, and Bystander Children's Intervention to Bullying. *Journal of Clinical Child and Adolescent Psychology*, *46*(3), 394–400. https://doi.org/10.1080/15374416.2015.1046180

Barr, A., Lane, T. & Nosenzo, D. (2018). On the social inappropriateness of discrimination. *Journal of Public Economics*, 164, 153–164. https://doi.org/10.1016/j.jpubeco.2018.06.004

Basford, T. E., Offermann, L. R. & Behrend, T. S. (2014). Do You See What I See? Perceptions of Gender Microaggressions in the

Workplace. *Psychology of Women Quarterly*, *38*(3), 340–349. https://doi.org/10.1177/0361684313511420

Bayles, M. D. (1986). Mid-level principles and justification. *NOMOS: Am. Soc'y Pol. Legal Phil.*, 28, 49–67.

Belavusau, U. (2012). Fighting Hate Speech through EU Law. *Amsterdam Law Forum*, 4(1), 20–35. https://hdl.handle.net/1814/20934

Bell, M. C. (2021). John Stuart Mill's Harm Principle and Free Speech: Expanding the Notion of Harm. *Utilitas*, 33(2), 162–179. https://doi.org/10.1017/S0953820820000229

Benesch, S. (2013). *Dangerous speech: A proposal to prevent group violence*. 2013-02-23. https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf

Bicchieri, C. & Mercier, H. (2014). Norms and Beliefs: How Change Occurs. In M. Xenitidou & B. Edmonds (Eds.), *The Complexity of Social Norms* (pp. 37–54). Springer International Publishing. https://doi.org/10.1007/978-3-319-05308-0

Bicchieri, C. (2016). *Norms in the wild*. Oxford University Press.

Blanchard, F. A., Crandall, C. S., Brigham, J. C. & Vaughn, L. A. (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology*, 79(6), 993–997. https://doi.org/10.1037/0021-9010.79.6.993

Boyne, S. M. (2010). Free Speech, Terrorism, and European Security: Defining and Defending the Political Community. *SSRN Electronic Journal,* 30(2), 417–483. https://doi.org/10.2139/ssrn.1591806

Camp, E. (2018). Insinuation, Common Ground, and the Conversational Record. In D. Fogal, D. W. Harris & M. Moss (Eds.), *New Work on Speech Acts* (Vol. 1, pp. 44–66). Oxford University Press. https://doi.org/10.1093/oso/9780198738831.003.0002

Cepollaro, B., Lepoutre, M. & Simpson, R. M. (2023). Counterspeech. *Philosophy Compass*, *18*(1). https://doi.org/10.1111/phc3.12890

Charles-Toussaint, G. C. & Crowson, H. M. (2010). Prejudice against International Students: The Role of Threat Perceptions and Authoritarian Dispositions in U.S. Students. *The Journal of Psychology*, *144*(5), 413–428. https://doi.org/10.1080/00223980.2010.496643

Chater, N. & Loewenstein, G. (2022). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences*, 1–60. https://doi.org/10.1017/S0140525X22002023

Christensen, J. F., Di Costa, S., Beck, B. & Haggard, P. (2019). I just lost it! Fear and anger reduce the sense of agency: a study using intentional binding. *Experimental Brain Research*, *237*(5), 1205–1212. https://doi.org/10.1007/s00221-018-5461-6

Christensen, R. (2022). *ordinal—Regression Models for Ordinal Data* (R package version 2022.11-16). https://cran.r-project.org/web/packages/ordinal/index.html

Cohen, M., Quintner, J. & van Rysewyk, S. (2018). Reconsidering the International Association for the Study of Pain definition of pain. *PAIN Reports*, 3(2), e634. https://doi.org/10.1097/PR9.0000000000000634

Considine, C. (2017). The Racialization of Islam in the United States: Islamophobia, Hate Crimes, and "Flying while Brown." *Religions*, 8(9), 165. https://doi.org/10.3390/rel8090165

Cook, W. L. & Sheppard, L. (2018). Not Doing Nothing: Third Parties' Cognitive Reactions to Mistreatment of Others. *Academy of Management Proceedings*, *2018*(1), 15718. https://doi.org/10.5465/AMBPP.2018.15718abstract

Curtis, V., Aunger, R. & Rabie, T. (2004). Evidence that disgust evolved to protect from risk of disease. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(suppl_4). https://doi.org/10.1098/rsbl.2003.0144

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F., Dreber, A., Wang, Y. & Costa, J. (2009). Accidental Outcomes Guide Punishment in a "Trembling Hand" Game.

PLoS ONE, 4(8), e6699. https://doi.org/10.1371/journal.pone.0006699

Cushman, F., Gray, K., Gaffey, A. & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, 12(1), 2–7. https://doi.org/10.1037/a0025071

Cushman, F. (2013). Action, Outcome, and Value. *Personality and Social Psychology Review*, 17(3), 273–292. https://doi.org/10.1177/1088868313495594

Czopp, A. M. & Monteith, M. J. (2003). Confronting Prejudice (Literally): Reactions to Confrontations of Racial and Gender Bias. *Personality and Social Psychology Bulletin*, 29(4), 532–544. https://doi.org/10.1177/0146167202250923

Darley, J. M. (2009). Morality in the Law: The Psychological Foundations of Citizens' Desires to Punish Transgressions. *Annual Review of Law and Social Science*, 5(1), 1–23. https://doi.org/10.1146/annurev.lawsocsci.4.110707.172335

de Araujo, E., Altay, S., Bor, A. & Mercier, H. (2020). Dominant jerks: People infer dominance from the utterance of challenging and offensive statements. *Social Psychological Bulletin*, 16(4). https://doi.org/10.32872/spb.6999

de Silva, A. & Simpson, R. M. (2022). Law as Counterspeech. *Ethical Theory and Moral Practice*. https://doi.org/10.1007/s10677-022-10335-3

Delgado, R. (1993). Words that Wound: A Tort Action for Racial Insults, Epithets and Name Calling. In *Words that Wound:*

Critical Race Theory, Assaultive Speech, and the First Amendment. Westview Press. https://scholarship.law.columbia.edu/books/287/

Delgado, R. (2013). The Harm in Hate Speech. *Law & Society Review*, 47(1), 232–233. https://doi.org/10.1111/lasr.12008

Dessel, A. B., Goodman, K. D. & Woodford, M. R. (2017). LGBT discrimination on campus and heterosexual bystanders: Understanding intentions to intervene. *Journal of Diversity in Higher Education*, 10(2), 101–116. https://doi.org/10.1037/dhe0000015

Dickter, C. L. & Newton, V. A. (2013). To confront or not to confront: non-targets' evaluations of and responses to racist comments. *Journal of Applied Social Psychology*, 43, E262–E275. https://doi.org/10.1111/jasp.12022

Dillon, K. P. & Bushman, B. J. (2015). Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context. *Computers in Human Behavior*, *45*(April), 144–150. https://doi.org/10.1016/j.chb.2014.12.009

Duckitt, J. & Sibley, C. G. (2017). The Dual Process Motivational Model of Ideology and Prejudice. In *The Cambridge Handbook of the Psychology of Prejudice* (pp. 188–221). Cambridge University Press. https://doi.org/10.1017/9781316161579.009

Dworkin, R. (1977). *Taking Rights Seriously*. Harvard University Press.

Eisenberger, N. I. (2015). Social Pain and the Brain: Controversies, Questions, and Where to Go from Here. *Annual Review of Psychology*, *66*(1), 601–629. https://doi.org/10.1146/annurev-psych-010213-115146

Ekins, E. (2017). *The State of free Speech and Tolerance in America*. https://www.cato.org/survey-reports/state-free-speech-tolerance-america

El Zein, M., Bahrami, B. & Hertwig, R. (2019). Shared responsibility in collective decisions. *Nature Human Behaviour*, 3(6), 554–559. https://doi.org/10.1038/s41562-019-0596-4

Eligon, J. (2018). Hate Crimes Increase for the Third Consecutive Year, F.B.I. Reports. *New York Times*. https://www.nytimes.com/2018/11/13/us/hate-crimes-fbi-2017.html

Fasoli, F., Carnaghi, A. & Paladino, M. P. (2015). Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Language Sciences*, 52, 98–107. https://doi.org/10.1016/j.langsci.2015.03.003

Fattoracci, E. S. M. & King, D. D. (2023). The Need for Understanding and Addressing Microaggressions in the Workplace. *Perspectives on Psychological Science*, *18*(4), 738–742. https://doi.org/10.1177/17456916221133825

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research*

Methods, 39(2), 175–191. https://doi.org/10.3758/BF03193146

Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fischer, P. & Greitemeyer, T. (2013). The positive bystander effect: Passive bystanders increase helping in situations with high expected negative consequences for the helper. *Journal of Social Psychology*, *153*(1), 1–5. https://doi.org/10.1080/00224545.2012.697931

Fraser, R. (2023). How to talk back: hate speech, misinformation, and the limits of salience. *Politics, Philosophy & Economics*, *22*(3), 315–335. https://doi.org/10.1177/1470594X231167593

Fumagalli, C. (2021). Counterspeech and Ordinary Citizens: How? When? *Political Theory*, 49(6), 1021–1047. https://doi.org/10.1177/0090591720984724

Gagliardone, I., Gal, D., Alves, T. & Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000233231

Galinsky, A. D., Wang, C. S., Whitson, J. A., Anicich, E. M., Hugenberg, K. & Bodenhausen, G. V. (2013). The Reappropriation of Stigmatizing Labels. *Psychological Science*, *24*(10), 2020–2029. https://doi.org/10.1177/0956797613482943

Gallup-Knight Foundation. (2020*). The First Amendment on campus 2020 report: College students' views of free expression*. https://knightfoundation.org/wp-content/uploads/2020/05/First-Amendment-on-Campus-2020.pdf

Gelber, K. (2012). Reconceptualizing Counterspeech in Hate Speech Policy (with a Focus on Australia). In *The Content and Context of Hate Speech* (pp. 198–216). Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.016

Gelber, K. & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324–341. https://doi.org/10.1080/13504630.2015.1128810

Gibson, J. L., Epstein, L. & Magarian, G. P. (2020). Taming Uncivil Discourse. *Political Psychology*, 41(2), 383–401. Advanced Online Publication. (2019). https://doi.org/10.1111/pops.12626

Gino, F., Moore, D. A. & Bazerman, M. H. (2008a). *No Harm, No Foul: The Outcome Bias in Ethical Judgments*. https://hbswk.hbs.edu/item/no-harm-no-foul-the-outcome-bias-in-ethical-judgments

Gino, F., Shu, L. L. & Bazerman, M. H. (2008b). Nameless Harmless = Blameless: When Seemingly Irrelevant Factors Influence Judgment of (Un)ethical Behavior. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1238661

Gino, F., Moore, D. A. & Bazerman, M. H. (2009). No Harm, No Foul: The Outcome Bias in Ethical Judgments. *SSRN Electronic Journal*, 52. https://doi.org/10.2139/ssrn.1099464

Githens-Mazer, J. & Lambert, R. (2010). *Islamophobia and Anti-Muslim Hate Crime: a London Case Study.* Exeter : University of Exeter, European Muslim Research Centre. https://unesdoc.unesco.org/ark:/48223/pf0000233231

Goldberg, S. C. (2010). The Epistemology of Silence. In *Social Epistemology* (pp. 243–261). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199577477.003.0012

Goldberg, S. C. (2020). *Conversational Pressure*. Oxford University Press.

Green, M. (2021). Speech Acts. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy Archive* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/speech-acts/

Greenberg, J. & Pyszczynski, T. (1985). The effect of an overheard ethnic slur on evaluations of the target: How to spread a social disease. *Journal of Experimental Social Psychology*, *21*(1), 61–72. https://doi.org/10.1016/0022-1031(85)90006-X

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in

Moral Judgment. *Science*, 293(5537), 2105 LP – 2108. https://doi.org/10.1126/science.1062872

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M. & Cohen, J. D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron*, 44(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8), 322–323. https://doi.org/10.1016/j.tics.2007.06.004

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E. & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001

Gulker, J. E., Mark, A. Y. & Monteith, M. J. (2013). Confronting prejudice: The who, what, and why of confrontation effectiveness. *Social Influence*, 8(4), 280–293. https://doi.org/10.1080/15534510.2012.736879

Hannikainen, I., Miller, R. & Cushman, F. (2014). If it feels bad to me, it's wrong for you: The role of emotions in evaluating harmful acts. *The Jury Expert*, 26. https://www.thejuryexpert.com/wp-content/uploads/1408/JuryExpert_1408_Wrong.pdf

Harvey, M. G., Buckley, M. R., Heames, J. T., Zinko, R., Brouer, R. L. & Ferris, G. R. (2007). A Bully as an Archetypal Destructive

Leader. *Journal of Leadership & Organizational Studies*, *14*(2), 117–129. https://doi.org/10.1177/1071791907308217

Henry, P. J., Butler, S. E. & Brandt, M. J. (2014). The influence of target group status on the perception of the offensiveness of group-based slurs. *Journal of Experimental Social Psychology*, 53, 185–192. https://doi.org/10.1016/j.jesp.2014.03.012

Heras-Escribano, M. (2019). *The Philosophy of Affordances*. Springer International Publishing. https://doi.org/10.1007/978-3-319-98830-6

Hester, N. & Gray, K. (2020). The Moral Psychology of Raceless, Genderless Strangers. *Perspectives on Psychological Science*, 15(2), 216–230. https://doi.org/10.1177/1745691619885840

Holm, G., Sahlström, F. & Zilliacus, H. (2018). Arts-Based Visual Research. In P. Leavy (Ed.), *Handbook of arts-based research* (pp. 311–335). Guilford Press.

Home Office. (2019). Hate crime, England and Wales, 2018 to 2019. *Home Office Statistical Bulletin 24 19. https://www.gov.uk/government/statistics/hate-crime-england-and-wales-2018-to-2019*

Home Office. (2021). Hate Crime, England and Wales, 2020 to 2021. *Home Office Statistical Bulletin 26 21. https://www.gov.uk/government/statistics/hate-crime-england-and-wales-2020-to-2021*

Hoover, J., Atari, M., Mostafazadeh Davani, A., Kennedy, B., Portillo-Wightman, G., Yeh, L. & Dehghani, M. (2021). Investigating the role of group-based morality in extreme behavioral expressions of prejudice. *Nature Communications*, *12*(1), 4585. https://doi.org/10.1038/s41467-021-24786-2

Hornsey, M. J. & Imani, A. (2004). Criticizing Groups from the Inside and the Outside: An Identity Perspective on the Intergroup Sensitivity Effect. *Personality and Social Psychology Bulletin*, 30(3), 365–383. https://doi.org/10.1177/0146167203261295

House, B. R. (2018). How do social norms influence prosocial development? *Current Opinion in Psychology*, 20, 87–91. https://doi.org/10.1016/j.copsyc.2017.08.011

Howard, J. W. (2021). Terror, Hate and the Demands of Counter-Speech. *British Journal of Political Science*, 51(3), 924–939. https://doi.org/10.1017/S000712341900053X

Hughes, J. (1984). Group Speech Acts. *Linguistics and Philosophy*, *7*(4), 379–395. https://www.jstor.org/stable/25001176

Janson, G. R., Carney, J. V., Hazler, R. J. & Oh, I. (2009). Bystanders' reactions to witnessing repetitive abuse experiences. *Journal of Counseling and Development*, *87*(3), 319–326. https://doi.org/10.1002/j.1556-6678.2009.tb00113.x

Janson, G. R. & Hazler, R. J. (2004). Trauma Reactions of Bystanders and Victims to Repetitive Abuse Experiences. *Violence and*

Victims, *19*(2), 239–255. https://doi.org/10.1891/vivi.19.2.239.64102

Jay, T. (2009). Do offensive words harm people? *Psychology, Public Policy, and Law*, *15*(2), 81–101. https://doi.org/10.1037/a0015646

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, 67(1), 85–92. https://doi.org/10.1093/biomet/67.1.85

Kawakami, K., Dunn, E., Karmali, F. & Dovidio, J. F. (2009). Mispredicting Affective and Behavioral Responses to Racism. *Science*, 323(5911), 276–278. https://doi.org/10.1126/science.1164951

Kazerooni, F., Taylor, S. H., Bazarova, N. N. & Whitlock, J. (2018). Cyberbullying Bystander Intervention: The Number of Offenders and Retweeting Predict Likelihood of Helping a Cyberbullying Victim. *Journal of Computer-Mediated Communication*, *23*(3), 146–162. https://doi.org/10.1093/jcmc/zmy005

Kellner, P. (2012). *Democracy on trial. What Voters Really Think of Parliment and our Politicians.* http://yougov.co.uk/news/2012/06/18/democracy-trial/

Keselman, H. J., Wilcox, R. R. & LIX, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology*, 40(4), 586–596. https://doi.org/10.1111/1469-8986.00060

Keshmirian, A., Hemmatian, B., Bahrami, B., Deroy, O. & Cushman, F. (2022). Diffusion of punishment in collective norm violations. *Scientific Reports*, 12(1), 15318. https://doi.org/10.1038/s41598-022-19156-x

Khanolainen, D. & Semenova, E. (2020). School Bullying Through Graphic Vignettes: Developing a New Arts-Based Method to Study a Sensitive Topic. *International Journal of Qualitative Methods*, 19. https://doi.org/10.1177/1609406920922765

Kim, Y. (2021). Understanding the Bystander Audience in Online Incivility Encounters: Conceptual Issues and Future Research Questions. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2934–2943. https://doi.org/10.24251/HICSS.2021.357

King, E. B., Dunleavy, D. G., Dunleavy, E. M., Jaffer, S., Morgan, W. B., Elder, K. & Graebner, R. (2011). Discrimination in the 21st century: Are science and the law aligned? *Psychology, Public Policy, and Law*, *17*(1), 54–75. https://doi.org/10.1037/a0021673

Kneer, M. & Skoczeń, I. (2023). Outcome effects, moral luck and the hindsight bias. *Cognition*, 232, 105258. https://doi.org/10.1016/j.cognition.2022.105258

Kukla, Q. R. (2023). Uptake and refusal. *Inquiry*, 1–27. https://doi.org/10.1080/0020174X.2023.2258207

Kumle, L., Võ, M. L.-H. & Draschkow, D. (2018). *Mixedpower: a library for estimating simulation-based power for mixed models in R*. (v1.0). Zenodo. https://doi.org/10.5281/zenodo.1341048

Langton, R. (1993). Speech Acts and Unspeakable Acts. *Philosophy & Public Affairs*, *22*(4), 293–330. www.jstor.org/stable/2265469

Langton, R. (2007). Disenfranchised Silence. In G. Brennan, R. Goodin, F. Jackson & M. Smith (Eds.), *Common Minds: Themes from the Philosophy of Philip Pettit* (pp. 199–214). Oxford University Press.

Langton, R. (2012). Beyond Belief: Pragmatics in Hate Speech and Pornography. In M. K. McGowan & I. Maitra (Eds.), *Speech and Harm: Controversies Over Free Speech* (pp. 144–164). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199236282.003.0004

Langton, R. (2018a). The Authority of Hate Speech. In J. Gardner, L. Green & B. Leiter (Eds.), *Oxford Studies in Philosophy of Law* (Vol. 3, pp. 132–152). Oxford University Press. https://doi.org/10.1093/oso/9780198828174.003.0004

Langton, R. (2018b). Blocking as Counter-Speech. In D. Fogal, D. W. Harris & M. Moss (Eds.), *New Work on Speech Acts* (pp. 1–36). Oxford University Press. https://doi.org/10.1093/oso/9780198738831.003.0006

Lawrence, F. M. (1993). The hate crimes/hate speech paradox : punishing bias crimes and protecting racist speech. *Notre Dame*

Law Review, 68, 673–721. http://hdl.handle.net/20.500.12389/20725

Leader Maynard, J. & Benesch, S. (2016). Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention. *Genocide Studies and Prevention*, *9*(3), 70–95. https://doi.org/10.5038/1911-9933.9.3.1317

Lench, H. C., Domsky, D., Smallman, R. & Darbor, K. E. (2015). Beliefs in moral luck: When and why blame hinges on luck. *British Journal of Psychology*, 106(2), 272–287. https://doi.org/10.1111/bjop.12072

Leonhard, L., Rueß, C., Obermaier, M. & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *Studies in Communication | Media*, *7*(4), 555–579. https://doi.org/10.5771/2192-4007-2018-4-555

Lepoutre, M. (2017). Hate Speech in Public Discourse. *Social Theory and Practice*, *43*(4), 851–883. https://doi.org/10.5840/soctheorpract201711125

Lepoutre, M. (2019). Hate Speech Laws: Expressive Power is not the Answer. *Legal Theory*, 25(4), 272–296. https://doi.org/10.1017/S135232522000004X

Lepoutre, M. (2021). *Democratic Speech in Divided Times*. OUP: Oxford University Press.

Lepoutre, M. (2023). Discursive optimism defended. *Politics, Philosophy & Economics*, *22*(3), 357–374. https://doi.org/10.1177/1470594X231179665

Lepoutre, M., Vilar-Lluch, S., Borg, E. & Hansen, N. (2023). What is Hate Speech? The Case for a Corpus Approach. *Criminal Law and Philosophy*. https://doi.org/10.1007/s11572-023-09675-7

Levin, B. & Reitzel, J. D. (2018). *Hate Crimes Rise in US Cities and Counties in Time of Division and Foreign Interference*. https://www.csusb.edu/sites/default/files/2018 Hate Final Report 5-14.pdf

Lewiński, M. (2021). Illocutionary pluralism. *Synthese*, *199*(3–4), 6687–6714. https://doi.org/10.1007/s11229-021-03087-7

Lutgen-Sandvik, P. & McDermott, V. (2008). The Constitution of Employee-Abusive Organizations: A Communication Flows Theory. Communication Theory, 18(2), 304–333. https://doi.org/10.1111/j.1468-2885.2008.00324.x

Lytle, R. D., Bratton, T. M. & Hudson, H. K. (2021). Bystander Apathy and Intervention in the Era of Social Media. *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*, 711–728. https://doi.org/10.1108/978-1-83982-848-520211052

Maitra, I. (2004). Silence and Responsibility. *Philosophical Perspectives*, 18(1), 189–208. https://doi.org/10.1111/j.1520-8583.2004.00025.x

Maitra, I. (2012). Subordinating Speech. In *Speech and Harm* (pp. 94–120). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199236282.003.0005

Maitra, I. & McGowan, M. K. (2012). Introduction and Overview. In I. Maitra & M. K. McGowan (Eds.), *Speech and Harm* (pp. 1–23). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199236282.001.0001

Matsuda, M. J., Lawrence III, C. R., Delgado, R. & Crenshaw, K. W. (1993). *Words That Wound* (1st ed.). Westview Press. https://scholarship.law.columbia.edu/books/287

May, L. (1990). Symposia Papers: Collective Inaction and Shared Responsibility. *Noûs*, *24*(2), 269–277. https://doi.org/10.2307/2215528

McDonald, L. (2021). Your word against mine: the power of uptake. *Synthese*, *199*(1–2), 3505–3526. https://doi.org/10.1007/s11229-020-02944-1

McDonald, L. (2022). Reimagining Illocutionary Force. *The Philosophical Quarterly*, *72*(4), 918–939. https://doi.org/10.1093/pq/pqab063

McGowan, M. K. (2018). Responding to Harmful Speech: The More Speech Response, Counter Speech, and the Complexity of Language Use. In C. R. Johnson (Ed.), *Voicing Dissent* (pp. 182–199). Routledge.

Mihajlova, E., Bacovska, J. & Sherkerdjiev, T. (2013). *Freedom of expression and hate speech*. OSCE, Mission to Skopje. https://vzs.ba/wp-content/uploads/2023/06/OSCE_-_Freedom_of_expression_and_hate_speech.pdf

Mill, J. S. (1859). *On liberty*. John W. Parker and Son.

Miller, R. & Cushman, F. (2013). Aversive for Me, Wrong for You: First-person Behavioral Aversions Underlie the Moral Condemnation of Harm. *Social and Personality Psychology Compass*, 7(10), 707–718. https://doi.org/10.1111/spc3.12066

Miller, R. M., Hannikainen, I. A. & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion*, 14(3), 573–587. https://doi.org/10.1037/a0035361

Mishna, F., Cook, C., Gadalla, T., Daciuk, J. & Solomon, S. (2010). Cyber bullying behaviors among middle and high school students. *American Journal of Orthopsychiatry*, *80*(3), 362–374. https://doi.org/10.1111/j.1939-0025.2010.01040.x

Monteith, M. J., Deneen, N. E. & Tooman, G. D. (1996). The effect of social norm activation on the expression of opinions concerning gay men and blacks. *Basic and Applied Social Psychology*, 18(3), 267–288. https://doi.org/10.1207/s15324834basp1803_2

Moreno, A. & Pérez-Navarro, E. (2021). Beyond the conversation: The pervasive danger of slurs. *Organon F*, *28*(3), 708–725. https://doi.org/10.31577/ORGF.2021.28311

Myers, W. & Lantz, B. (2020). Reporting Racist Hate Crime Victimization to the Police in the United States and the United Kingdom: A Cross-National Comparison. *The British Journal of Criminology*, 60(4), 1034–1055. https://doi.org/10.1093/bjc/azaa008

Namie, G. & Lutgen-Sandvik, P. E. (2010). Active and Passive Accomplices: The Communal Character of Workplace Bullying. *International Journal of Communication*, *4*, 343–373. https://ijoc.org/index.php/ijoc/article/view/589

Naughton, K. A. et al. (2017). *Speaking Freely: What Students Think about Expression at American Colleges*. https://www.thefire.org/research-learn/student-attitudes-free-speech-survey

Ng, K., Niven, K. & Notelaers, G. (2022). Does Bystander Behavior Make a Difference? How Passive and Active Bystanders in the Group Moderate the Effects of Bullying Exposure. *Journal of Occupational Health Psychology*, *27*(1), 119–135. https://doi.org/10.1037/ocp0000296

Nielsen, L. B. (2009). *License to Harass*. Princeton University Press. https://doi.org/10.1515/9781400826292

Nielsen, L. B. (2012). Power in Public. In I. Maitra & M. K. McGowan (Eds.), *Speech and Harm* (pp. 148–173). Oxford University

Press. https://doi.org/10.1093/acprof:oso/9780199236282.003.0007

Opp, K.-D. (2001). How do norms emerge? An outline of a theory. *Mind & Society*, 2(1), 101–128. https://doi.org/10.1007/bf02512077

Opp, K. D. (2002). When do norms emerge by human design and when by the unintended consequences of human action? The example of the no-smoking norm. *Rationality and Society*, 14(2), 131–158. https://doi.org/10.1177/1043463102014002001

OSCE Office for Democratic Institutions and Human Rights. (2014). *Hate Crime Data-Collection and Monitoring Mechanisms*. OSCE ODIHR. https://www.osce.org/odihr/datacollectionguide

Palan, S. & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Parkinson, B. & Manstead, A. S. R. (1993). Making sense of emotion in stories and social life. *Cognition & Emotion*, 7(3–4), 295–323. https://doi.org/10.1080/02699939308409191

Perry, B. & Alvi, S. (2012). 'We are all vulnerable.' *International Review of Victimology*, 18(1), 57–71. https://doi.org/10.1177/0269758011422475

Petersson, B. (2008). Collective omissions and responsibility. *Philosophical Papers*, *37*(2), 243–261. https://doi.org/10.1080/05568640809485221

Petersen, T. S. (2016). No Offense! On the Offense Principle and Some New Challenges. *Criminal Law and Philosophy*, 10(2), 355–365. https://doi.org/10.1007/s11572-014-9333-2

Popa-Wyatt, M. & Wyatt, J. L. (2018). Slurs, roles and power. *Philosophical Studies*, *175*(11), 2879–2906. https://doi.org/10.1007/s11098-017-0986-2

Popa-Wyatt, M. (2021). Slurring Speech and Social Norms. In *The Social Institution of Discursive Norms* (p. 10). Routledge. https://doi.org/10.4324/9781003047483-15

Prochownik, K. & Cushman, F. A. (2019). Outcomes speak louder than actions? Testing a challenge to the two-process model of moral judgment. *CogSci*, 2506–2613. https://api.semanticscholar.org/CorpusID:201916172

Przepiorka, W., Szekely, A., Andrighetto, G., Diekmann, A. & Tummolini, L. (2022). How Norms Emerge from Conventions (and Change). *Socius*, 8. https://doi.org/10.1177/23780231221124556

Raja, S. N., Carr, D. B., Cohen, M., Finnerup, N. B., Flor, H., Gibson, S., Keefe, F. J., Mogil, J. S., Ringkamp, M., Sluka, K. A., Song, X.-J., Stevens, B., Sullivan, M. D., Tutelman, P. R., Ushida, T. & Vader, K. (2020). The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. *Pain*, 161(9), 1976–1982. https://doi.org/10.1097/j.pain.0000000000001939

Redish, M. H. (1984). *Freedom of Expression: A Critical Analysis*. Michie Co.

Rosenfeld, M. (2003). Hate Speech in Constitutional Jurisprudence: A Comparative Analysis. *Cardozo L Rev*, 24(4), 1523–1567. https://doi.org/10.1017/CBO9781139042871.018

Rovira, A., Southern, R., Swapp, D., Campbell, C., Zhang, J. J., Levine, M. & Slater, M. (2021). Bystander Affiliation Influences Intervention Behavior: A Virtual Reality Study. *SAGE Open*, *11*(3). https://doi.org/10.1177/21582440211040076

Rungtusanatham, M., Wallin, C. & Eckerd, S. (2011). The Vignette in a Scenario-Based Role-Playing Experiment. *Journal of Supply Chain Management*, 47(3), 9–16. https://doi.org/10.1111/j.1745-493X.2011.03232.x

Rupert, D. J., Poehlman, J. A., Hayes, J. J., Ray, S. E. & Moultrie, R. R. (2017). Virtual Versus In-Person Focus Groups: Comparison of Costs, Recruitment, and Participant Logistics. *Journal of Medical Internet Research*, 19(3), e80. https://doi.org/10.2196/jmir.6980

Sbisà, M. (1984). On illocutionary types. *Journal of Pragmatics*, *8*(1), 93–112. https://doi.org/10.1016/0378-2166(84)90066-3

Sbisà, M. (2001). Illocutionary force and degrees of strength in language use. *Journal of Pragmatics*, *33*(12), 1791–1814. https://doi.org/10.1016/S0378-2166(00)00060-6

Scanlon, T. (1972). A Theory of Freedom of Expression. *Philosophy & Public Affairs*, 1(2), 204–226. http://www.jstor.org/stable/2264971

Schauer, F. F. (1982). *Free Speech: A Philosophical Enquiry*. Cambridge University Press.

Schauer, F. (1993). The Phenomenology of Speech and Harm. *Ethics*, 103(4), 635–653. http://www.jstor.org/stable/2381631

Schauer, F. (2015). On the distinction between speech and action. *Emory LJ*, 65(427). https://scholarlycommons.law.emory.edu/elj/vol65/iss2/6/

Schein, C. & Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review*, 22(1), 32–70. https://doi.org/10.1177/1088868317698288

Schein, C. (2020). The Importance of Context in Moral Judgments. *Perspectives on Psychological Science*, 15(2), 207–215. https://doi.org/10.1177/1745691620904083

Schmader, T., Croft, A., Scarnier, M., Lickel, B. & Mendes, W. B. (2012). Implicit and explicit emotional reactions to witnessing prejudice. *Group Processes & Intergroup Relations*, *15*(3), 379–392. https://doi.org/10.1177/1368430211426163

Schmid, H. B. (2014a). Expressing Group Attitudes: On First Person Plural Authority. *Erkenntnis*, *79*(S9), 1685–1701. https://doi.org/10.1007/s10670-014-9635-8

Schmid, H. B. (2014b). Plural self-awareness. *Phenomenology and the Cognitive Sciences*, 13(1), 7–24. https://doi.org/10.1007/s11097-013-9317-z

Schmid, H. B. (2018a). The subject of "We intend." *Phenomenology and the Cognitive Sciences*, *17*(2), 231–243. https://doi.org/10.1007/s11097-017-9501-7

Schmid, H. B. (2018b). Collective Responsibilities of Random Collections: Plural Self-Awareness among Strangers. *Journal of Social Philosophy*, *49*(1), 91–105. https://doi.org/10.1111/josp.12229

Schmid, H. B. (2023). *We, Together*. Oxford University Press.

Schmitz, M. & Townsend, L. (2020). Introduction to Special Issue on 'Group Speech Acts.' *Language & Communication*, *72*, 53–55. https://doi.org/10.1016/j.langcom.2020.03.001

Searle, J. R. (1979). *Expression and Meaning*. Cambridge University Press. https://doi.org/10.1017/CBO9780511609213

Skarlicki, D. P. & Kulik, C. T. (2004). Third-Party Reactions to Employee (Mis)Treatment: A Justice Perspective. *Research in Organizational Behavior*, *26*, 183–229. https://doi.org/10.1016/S0191-3085(04)26005-1

Skarlicki, D. P., O'Reilly, J. & Kulik, C. T. (2015). The Third-Party Perspective of (In)justice. In R. S. Cropanzano & M. L. Ambrose (Eds.), *The Oxford Handbook of Justice in the Workplace* (pp. 235–256). Oxford University Press.

https://doi.org/10.1093/oxfordhb/9780199981410.001.0001

Soral, W., Bilewicz, M. & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Spanish Ministry of Interior. (2019). *Action Plan to Combat Hate Crimes*. https://www.interior.gob.es/opencms/pdf/servicios-al-ciudadano/Delitos-de-odio/descargas/PLAN-DE-ACCION-DE-LUCHA-CONTRA-LOS-DELITOS-DE-ODIO-english-version.pdf

Stanley, J. (2015). *How Propaganda Works*. Princeton University Press.

Stevens, J.P. (2010) *Opinion: Citizens United v. FEC, 558 U.S. 310*, https://www.law.cornell.edu/supct/html/08-205.ZX.html

Sullivan, G. M. & Artino, A. R. (2013). Analyzing and Interpreting Data From Likert-Type Scales. *Journal of Graduate Medical Education*, 5(4), 541–542. https://doi.org/10.4300/JGME-5-4-18

Swim, J. K. & Hyers, L. L. (1999). Excuse Me—What Did You Just Say?!: Women's Public and Private Responses to Sexist Remarks. *Journal of Experimental Social Psychology*, *35*(1), 68–88. https://doi.org/10.1006/jesp.1998.1370

Swim, J. K., Hyers, L. L., Cohen, L. L. & Ferguson, M. J. (2001). Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of*

Social Issues, *57*(1), 31–53. https://doi.org/10.1111/0022-4537.00200

Swim, J. K., Hyers, L. L., Cohen, L. L., Fitzgerald, D. C. & Bylsma, W. H. (2003). African American college students' experiences with everyday racism: Characteristics of and responses to these incidents. *Journal of Black Psychology*, *29*(1), 38–67. https://doi.org/10.1177/0095798402239228

Tabachnick, B. G. & Fidell, L. S. (2012). Principal Components and Factor Analysis. In *Using multivariate statistics* (Vol. 6, pp. 612–680). Pearson London.

Tännsjö, T. (2007). The Myth of Innocence: On Collective Responsibility and Collective Punishment. *Philosophical Papers*, *36*(2), 295–314. https://doi.org/10.1080/05568640709485203

Taylor, S. H., DiFranzo, D., Choi, Y. H., Sannon, S. & Bazarova, N. N. (2019). Accountability and Empathy by Design. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–26. https://doi.org/10.1145/3359220

Lutgen-Sandvik, P. & McDermott, V. (2008). The Constitution of Employee-Abusive Organizations: A Communication Flows Theory. Communication Theory, 18(2), 304–333. https://doi.org/10.1111/j.1468-2885.2008.00324.x

UK Home Office. (2016). *Action Against Hate: The UK Government's Plan for Tackling Hate Crime*.

https://www.gov.uk/government/publications/hate-crime-action-plan-2016

United States Department of Justice. (2018). *Hate Crime Statistics, 2017*. https://ucr.fbi.gov/hate-crime/2017

Urschler, D. F., Fischer, J., Kastenmüller, A. & Fischer, P. (2015). Bystander Effect. *Oxford Bibliographies*, *July*. https://doi.org/10.1093/obo/9780199828340-0172

Villacorta, Pablo, J. (2017). The welchADF Package for Robust Hypothesis Testing in Unbalanced Multivariate Mixed Models with Heteroscedastic and Non-normal Data. *The R Journal*, 9(2), 309. https://doi.org/10.32614/RJ-2017-049

Wabnitz, P., Martens, U. & Neuner, F. (2012). Cortical reactions to verbal abuse. *NeuroReport*, 23(13), 774–779. https://doi.org/10.1097/WNR.0b013e328356f7a6

Waldron, J. (2012). *The Harm in Hate Speech*. Harvard University Press. https://www.jstor.org/stable/j.ctt2jbrjd

Walters, M. A. (2014a). *Hate Crime and Restorative Justice*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199684496.001.0001

Walters, M. A. (2014b). The Harms of Hate Crime: From Structural Disadvantage to Individual Identity. In *Hate Crime and Restorative Justice: Exploring Causes, Repairing Harms* (pp. 62–90). Oxford University Press.

https://doi.org/10.1093/acprof:oso/9780199684496.003.0003

Wang, C. S., Whitson, J. A., Anicich, E. M., Kray, L. J. & Galinsky, A. D. (2017). Challenge Your Stigma. *Current Directions in Psychological Science*, *26*(1), 75–80. https://doi.org/10.1177/0963721416676578

Weinberg, J. D. & Nielsen, L. B. (2017). What is Sexual Harassment? An Empirical Study of Perceptions of Ordinary People and Judges. *Saint Louis University Public Law Review*, 36(1), Article 6. https://scholarship.law.slu.edu/plr/vol36/iss1/6/

Welch, B. L. (1951). On The Comparison of Several Mean Values: An Alternative Approach. *Biometrika*, 38(3–4), 330–336. https://doi.org/10.1093/biomet/38.3-4.330

Wenik, J. (1985). Forcing the Bystander to Get Involved: A Case for a Statute Requiring Witnesses to Report Crime. *The Yale Law Journal*, *94*(7), 1787. https://doi.org/10.2307/796222

Wike, R. & Simmons, K. (2015). Global support for principle of free expression, but opposition to some forms of speech. *Pew Research Center*. http://www.pewglobal.org/2015/11/18/global-support-for-principle-of-free-expression-but-opposition-to-some-forms-of-speech/

Witek, M. (2013). *How to Establish Authority with Words*. Logic, Methodology and Philosophy of Science at Warsaw University, 2(2011), 145–157. https://philpapers.org/rec/WITHTE

Whitson, J., Anicich, E. M., Wang, C. S. & Galinsky, A. D. (2017). Navigating Stigma and Group Conflict: Group Identification as a Cause and Consequence of Self-Labeling. *Negotiation and Conflict Management Research*, *10*(2), 88–106. https://doi.org/10.1111/ncmr.12094

Wong, R. Y. M., Cheung, C. M. K., Xiao, B. & Thatcher, J. B. (2021). Standing Up or Standing By: Understanding Bystanders' Proactive Reporting Responses to Social Media Harassment. *Information Systems Research*, 32(2), 561–581. https://doi.org/10.1287/isre.2020.0983

Young, L., Cushman, F., Hauser, M. & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104

Zapata, J. & Deroy, O. (2023). Ordinary citizens are more severe towards verbal than nonverbal hate-motivated incidents with identical consequences. *Scientific Reports*, 13(1), 1–14. https://doi.org/10.1038/s41598-023-33892-8

Zapata, J., Sulik, J., von Wulffen, C. & Deroy, O. (2024). Bystanders' collective responses set the norm against hate speech. Humanities and Social Sciences Communications, 11(1), 335. https://doi.org/10.1057/s41599-024-02761-8

Zitek, E. M. & Hebl, M. R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of*

Experimental Social Psychology, 43(6), 867–876. https://doi.org/10.1016/j.jesp.2006.10.010