iscte

INSTITUTO
UNIVERSITÁRIO
DE LISBOA

“Your post has been deleted”: Effects of ideological beliefs, threat, and social norm on support of hate speech prohibition

Amanda Salvador de Andrade

Master in Psychology of Intercultural Relations

Supervisor:
P.h.D. Sven Waldzus, Full Professor,
CIS-IUL - ISCTE – University Institute of Lisbon

Outubro, 2022.

CIÊNCIAS SOCIAIS
E HUMANAS

“Your post has been deleted”: Effects of ideological beliefs, threat, and social norm on support of hate speech prohibition

Amanda Salvador de Andrade

Master in Psychology of Intercultural Relations

Supervisor:
P.h.D. Sven Waldzus, Full Professor,
CIS-IUL - ISCTE – University Institute of Lisbon

Outubro, 2022.

À memória da minha avó Cipriana Paz Martinez

Acknowledgments

To Doctor Professor Sven Waldzus for the guidance and sharing of knowledge. All reflections were essential for the conclusion of this investigation, and your attention, availability, and trust.

To my parents and brother, because without them the dream would not be possible. There are no words that describe the deep love, affection, and thanks for all the support throughout my career.

To Giovanna for respecting all my highs and lows during the preparation of the thesis, for her tireless patience, for her support, and especially for believing that I was capable.

To all my friends in Lisbon, Augusto, Carol, Conrado, Daniel, Flávia, Gonçalo, Isabella, Isadora, Leticia, and Sofia. I thank you for making living abroad easier and more fun.

A big thanks to all of you.

Resumo

O presente estudo examinou o efeito do autoritarismo no apoio à proibição do discurso de ódio online. Um estudo online (n=293) investigando se o autoritarismo de direita (V.I) irá prever positivamente o apoio à proibição do discurso de ódio (V.D). O estudo complementa as pesquisas existentes, examinando se a normatividade e a percepção de ameaça do discurso de ódio modera a relação entre o autoritarismo de direita e a proibição do discurso de ódio. Comparamos discurso de ódio censurado versus discurso de ódio não censurado (dentro do assunto), a manipulação da ameaça foi aleatoriamente designados para uma das três condições: ameaça à imigrante, ameaça aos LGBT e condição de controle (entre-sujeito). Nossos resultados mostram que os autoritários, independentemente da normatividade, não impedem o discurso de ódio mais do que os não autoritários e que a ameaça percebida aumenta a intenção de permitir e a intenção para reportar o discurso de ódio em autoritários, independentemente da normatividade do discurso de ódio

Palavras-chave:
Psicologia social, discurso de ódio, discurso de ódio online, LGBT, imigrante, Portugal.
Códigos PsycINFO:
3100 Personality Psychology
2750 Mass Media Communications
2900 Social Processes & Social Issues

Abstract

The present study examined the effect of authoritarianism on the support of online Hate Speech Prohibition. An online study (n=293) investigating if right-wing authoritarianism (I.V) will positively predict support of the prohibition of hate speech (D.V). The study adds to the existing research by examining whether the normativity and perceived threat of hate speech moderates the relation between right-wing authoritarianism and hate speech prohibition. We compare censored hate speech vs non-censored hate speech (within-subject). Threat manipulation were randomly assigned to one of the three conditions: immigrant threat, LGBT threat, and control condition (between-subject). Our results show that authoritarians, regardless of normativity, do not prevent hate speech more than non-authoritarians do and that perceived threat increases the willingness to permit and the willingness to report hate speech in authoritarians, independently from normativity of hate speech

Keywords:
Social psychology, hate speech, online hate speech, LGBT, immigrant, Portugal.
PsycINFO Codes:
3100 Personality Psychology
2750 Mass Media Communications
2900 Social Processes & Social Issues

Index

Introduction

As an instrument of communication, a way of sharing information, photo, and video, social networks are important ways of reproducing real situations and connecting people in a virtual environment. In this context, social phenomena are also observed in online settings, including the phenomenon of hate speech. This phenomenon has always been present in society but is now replicated and amplified on the networks.

The role of social media in the public debate is significant. Social debate in social media can reflect real-life situations, as well as offline debate can lead to online situations. For example, during the COVID-19 lockdown, the hashtag movement #JeNeSuisPasUnVirus (I am not a virus) spread across Europe and the world, reflecting a counter-speak response from an Asian minority in Europe against prejudice suffered during COVID-19 (Bayer & Bárd, 2020). Also, in 2018, the United Nations (UN) accused Facebook of having played an important role in the genocide in Myanmar, by helping to spread hate speech online. The UN investigated the social network providor's slow response to ban the fast spread of Islamophobic hate speech against the ongoing persecutions and killings of the Muslim Rohingya people in Myanmar. (OHCHR, 2018).

Recently, in 2021, Casa do Brasil de Lisboa (CBL) in a project collaboration with #MigraMyths – *Desmistificando a Imigração* conducted an online survey about hate speech against immigrants in Portugal. The study was conducted in Portugal with people from Portugal, South America, Southern Europe, Africa, and Pakistan, with most of the participants from Brazil (66.0%). The research has shown that 75.4% of participants have suffered some kind of hate speech based on prejudice and stereotypes about immigration or for being an immigrant in Portugal. Additionally, when participants were asked where they perceived most of the dissemination of hate speech, the answers were: 32.4% on the internet (social media platforms, like Facebook, Instagram, and Twitter), 20.9% in public services (Government institutions), and 19.6% in education institutions (Schools and Universities) (Casa do Brasil de Lisboa, 2021).

These are examples of the importance of reflecting on the complexity of online hate speech. Consequently, this work, through experimental research, seeks to understand what motivations are involved in responses toward hate speech on social networks. Different from other studies that assumed that conformity to norms is the cause of prejudice (Adorno et al., 1950; Allport, 1954), Bilewicz et al.’s (2015) research shows that individuals who have a strong belief in social norms, measured by the Right-Wing Authoritarianism (RWA) scale, are more likely to oppose and prohibit hate speech on social media once derogating expressions would not be accepted socially (Bilewicz et al., 2015; Bilewicz & Soral, 2020). The current study will add to the existing research by examining the potential moderation of this relation between RWA and HS prohibition by threat and normativity.

The first chapter presents and discusses hate speech definitions and reflects on the lack of consensus among organizations and nations about the concept. The subsequent sections present the theoretical framework, methods, and results. Finally, Chapter 4 presents the discussion.

CHAPTER 1 –

Literature Review

1.1. Hate Speech Definition

In recent years, the internet has become an important tool for everyday life, which allows us to establish connections with people and the world. Nowadays, the internet and social media platforms have given hate speech new forms and dimensions, although hate speech is not a recent phenomenon. Historically, mass media, such as newspapers and radio, had shaped the use of these expressions. For instance, in Nazi propaganda, dehumanizing terms ("rats" or "vermin") to address Jewish people were widely used and disseminated by mass media (Carlson, 2021).

More importantly, dealing with hate speech in society brings problems that range from the difficulty to find a consensual definition of the concept to its regulation, in conjunction with the tension between the concerns with hate speech and the guarantee of freedom of expression. As many researchers have reported, "defining hate speech poses challenges because of differences in social and legal contexts" (Gonçalves et al., 2021, p.5). The difficulties in defining this phenomenon, as well as its restriction, associated with the current sociopolitical scenario, have provided a rich environment for the evolution of its discussion.

Because of the complexity of the phenomenon, different sources and institutions define the term hate speech in a distinct way. For example, a legal accepted definition worldwide is the United Nations Office of the High Commissioner for Human Rights (OHCHR), which defined hate speech as "Advocacy of discriminatory hatred that constitutes incitement to hostility, discrimination, or violence" (Article 19 (Organization), 2012, p.8). We can recognize these incitements to have the intent to call for violence or discrimination. Although, it is a widely used and commonly accepted definition, is still too narrow and does not specify any target group, also, expressions such as insults or negative stereotyping, are not mentioned in the OHCHR definition.

Regarding the European context, The European Commission Against Racism and Intolerance -ECRI (2016) defined hate speech as follows:

> The use of one or more particular forms of expression—namely, the advocacy, promotion, or incitement of the denigration, hatred, or vilification of a person or group of persons, as well any harassment, insult, negative stereotyping, stigmatization, or

> threat of such person or persons and any justification of all these forms of expression – that is based on a non-exhaustive list of personal characteristics or status that includes “race”, color, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity, and sexual orientation. (p. 16)

Compared to the OHCHR definition, the ECRI definition focuses on disadvantaged social groups based on their characteristics or immutable characteristics. Also, the ECRI definition did not include “expressions that merely distress, hurt, or offend because hate speech is much more than mere dislike or bias, and it tends to be discriminatory, abusive, and hostile in nature” (Gonçalves et al., 2021, p.5)

Regarding activist organizations, for example, the NGO, Intervenção Lésbica, Gay, Bisexual e Trans e Intersex (ILGA), defined hate speech as “public expression which spread, incites, promote, or justifies hatred, discrimination, or hostility towards a specific group. They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups.”(ILGA Europe, 2021, n.d). The ILGA definition called attention to the dangers associated with that hate speech, especially the fact that hate speech can contribute to intolerance, which can make attacks more probable against a group.

Carlson (2021) defends that hate speech is an expression (verbal, symbolic, and imagery) that represents a “structural phenomenon in which those in power use verbal assaults and offensive imagery to maintain their preferred position in the existing social order” (Carlson, 2021, p.6). Carlson's definition (2021) shows the complexity and diversity of the concept of hate speech by considering power relations as an important variable for hate speech. However, the dilemma with this definition is that people who do not have social power can also commit hate speech, as well as people with social power can be the target of hate.

This author highlights not only what defines hate but also what is not considered hate speech. According to Carlson (2021), we cannot reduce hate speech to offensive speech. Therefore, affirming that a person has negative feelings about someone or hates a personality trait – does not constitute hate speech, as the author highlights:

> Hate speech is not synonymous with offensive speech. Words or images that someone finds upsetting or hurtful do not meet legal or even colloquial definitions of hate speech. Saying that you don’t like someone, their personality, or their politics, does not constitute hate speech. In order to be considered hate speech, expression must directly attack a person’s immutable identity characteristics such as race, gender, or sexual orientation (p.6)

This means a hateful statement against a group of journalists, a left-wing politician, or a group of anti-vaccination people may not constitute an attack of hate speech because those groups are not considered protected groups or groups based on an individual trait. It is worth highlighting that this understanding of hate speech is shared by most legal and social institutions and researchers. On the other hand, there are other lines of understanding that comprehend any group as potentially able or likely to suffer hate speech, depending on the social context.

Another complex facet of hate speech is that some expressions are based on religious beliefs or ideological beliefs, which can sometimes blur the fine line between hate speech and expressions of beliefs that are covered by the right to Freedom of Speech. In other words, it is not always easy to tell the difference between criticism or opinion (guaranteed by freedom of expression) and a speech that "hurts" others and encourages prejudice (hate speech). To address this issue, Bilewicz and Soral (2020) made an important difference between derogatory language (hate-speech) and non-derogatory forms of intergroup criticism. The authors elucidate an example of anti-Semitism hate speech ("derogatory discourse") and criticism of Israel ("non-derogatory form"). According to authors if the criticism of Israel is oriented by theories based on concerns with human rights and political considerations will hardly be considered hate speech. On the other hand, criticism of Israel based only in prejudice believes against Jews might be considered hateful expressions. This example explains the level of interpretation of hate speech depending on the social context.

Like Carlson (2021), other authors (Kaplin, 2016; Tynes et al., 2013) argue that hate speech can be expressed by symbolic acts with negative connotations. Thus, symbolic acts may appear as a joke or a cartoon in a student group chat (Kaplin, 2016) or as a Nazi or Ku Klux Klan parade, hate website, or cross-burnings (Boeckmann & Turpin-Petrosino, 2002). Subsequently, those researchers have focused on the harm hate-speech causes to the victim. For them, the major concern about hate speech is the consequences caused to the target group rather than the hateful content itself or the hateful motivation. (Barendt, 2019; Kaplin, 2016)

In Portugal, the Penal Code punishes acts of hate crimes, including hate speech, in different instances. As described in article 240º - *Discriminação e incitamento ao ódio e à violência*, whoever founds or sets up or provides assistance to an organization or carries out or participates in organized propaganda activities that incite or encourage discrimination, hatred, or violence against a person or group of persons because of their race, color, ethnic or national origin, ancestry, religion, sex, sexual orientation, gender identity or physical or mental disability shall be punished with imprisonment from 1 to 8 years. Moreover, whoever, publicly, by any means intended for dissemination, namely through apologia, denial or gross trivialization of crimes of genocide, war or against peace and humanity, provokes acts of violence against, defames, insults, threatens or incites violence or hatred towards persons or a group of persons because of their race, color, ethnic or national origin, ancestry, religion, sex, sexual orientation, gender identity or physical or psychological disability is punishable by imprisonment from 6 months to 5 years.[1]

As one may observe, law, researchers, and activist organizations have focused on different aspects of hate speech in their definitions. Some definitions center on the hate of minority groups or a list of specific groups (Carlson, 2021; ILGA Europe, n.d; ECRI, 2016). On the contrary, some accept that hate speech can target any group once it incites violence and discrimination (Article 19 (Organization), 2012).

In sum, it is possible to differentiate, in this literature review, five theoretical approaches to defining hate speech: 1) intention to cause or advocate harm or to discriminate against someone; 2) advocacy, promotion, or incitement of the denigration, hatred, vilification, harassment, negative stereotyping, and threat to groups based in characteristics; 3) expression of hostility and discrimination that may provoke attacks on minority groups; 4) the perception and potential damage caused by negative expressions and 5) expressions with the purpose to maintain a power position.

Private companies such as social media platforms (Facebook, Twitter, Instagram, GAB) and websites (Google, Journals websites) usually have their own definition of hate speech and interpretation of Free Speech. Thus, based on established policies from each platform, a speech can be forbidden or allowed to stay online, depending on the platform policies (Gagliardone, 2015). In addition, the platform norms include the possibility of self-regulation, which means

[1] Portuguese original: “1 - Quem:
a) Fundar ou constituir organização ou desenvolver atividades de propaganda organizada que incitem à discriminação, ao ódio ou à violência contra pessoa ou grupo de pessoas por causa da sua raça, cor, origem étnica ou nacional, ascendência, religião, sexo, orientação sexual, identidade de género ou deficiência física ou psíquica, ou que a encorajem; ou
b) Participar na organização ou nas atividades referidas na alínea anterior ou lhes prestar assistência, incluindo o seu financiamento;
é punido com pena de prisão de 1 a 8 anos.
2 - Quem, publicamente, por qualquer meio destinado a divulgação, nomeadamente através da apologia, negação ou banalização grosseira de crimes de genocídio, guerra ou contra a paz e a humanidade:
a) Provocar atos de violência contra pessoa ou grupo de pessoas por causa da sua raça, cor, origem étnica ou nacional, ascendência, religião, sexo, orientação sexual, identidade de género ou deficiência física ou psíquica;
b) Difamar ou injuriar pessoa ou grupo de pessoas por causa da sua raça, cor, origem étnica ou nacional, ascendência, religião, sexo, orientação sexual, identidade de género ou deficiência física ou psíquica;
c) Ameaçar pessoa ou grupo de pessoas por causa da sua raça, cor, origem étnica ou nacional, ascendência, religião, sexo, orientação sexual, identidade de género ou deficiência física ou psíquica; ou
d) Incitar à violência ou ao ódio contra pessoa ou grupo de pessoas por causa da sua raça, cor, origem étnica ou nacional, ascendência, religião, sexo, orientação sexual, identidade de género ou deficiência física ou psíquica;
é punido com pena de prisão de 6 meses a 5 anos.” (Decreto-Lei n.º 400/82, de 23 de setembro na 7 versão correspondente ao Decreto-Lei n.º 48/95)

the platform users can report a specific comment or imagery on the platform. These interventions aim to change people's online behavior and encourage individuals or groups to conform to established social platform norms and not to allow hateful content. Platform users themselves can report comments they judge as inappropriate. These comments then stay on the online platform, until the final decision of whether they will be deleted or flagged is taken, depending on the platform's policies. Therefore, the majority of social platforms have content moderation, being done either automatically by algorithms, content moderator employees, or by the users themselves that report inappropriate comments or posts. It is noteworthy that the defenders of Freedom of Speech have gradually migrated to special platforms, such as GAB, Reddit*,* and Rumble where there is no -–or very little -–content moderation.

Taking into account the complexity of the hate speech concept, for the current study we understand hate speech as (1) an expression that targets social groups based on specific characteristics and (2) a negative expression that intends to cause serious discrimination and harm to the target group.

1.2. Hate Speech in Social Psychology Studies

Social psychology studies on Labelling Theory have been supporting Allport's (1954) claim that "Antilocution" – speaking against – prejudice speech, such as hate speech, can lead to outgroup avoidance and discrimination and in more extreme cases physical violence (Bilewicz & Soral, 2020; Windisch et al., 2020). For instance, the use of slurs is a common form of hate speech. Research on labelling theory has studied the difference between ways to address the LGBT Community, for instance, the difference between the use of slur ("fag" or "fairy") or labels such as "gay" or "homosexual" (Carnaghi & Maass 2007).

Recent social psychological research has looked at the perception of social norms regarding hate speech instead of its effects. The importance of social norms is that they "can be seen as one of the factors preventing the spread of hate speech" (Bilewicz et al., 2015, p.13) due to the understanding of hate speech as a norm violation. More precisely, research has shown that individuals perceive hate speech as a case of norm violation in modern societies (Bilewicz et al., 2015). Different from other studies that assumed that conformity to norms is the cause of prejudice (Adorno eat, 1950; Allport, 1954), Bilewicz et al.'s (2015) research shows that individuals who have a strong belief in social norms, measured by the right-wing authoritarianism (RWA) scale, are more likely to oppose and prohibit hate speech on social media once derogating expressions would not be accepted socially (Bilewicz et al., 2015; Bilewicz & Soral, 2020).

Thus, Bilewicz et al. (2015) found that Authoritarianism can be a protector against hate speech online, despite the positive correlation of prejudiced attitudes with authoritarianism. This result might be due to authoritarians' understanding of hate speech expressions as socially deviant. According to Bilewicz et al.'s (2015) findings, people with a higher score on the RWA scale can be expected to be more likely to support the prohibition of Hate Speech.

Nevertheless, individuals scoring high on RWA were also more likely to score high in social distance from minorities, which is a subtle way of expressing prejudice (Bilewicz et al., 2015). In the paper, Bilewicz et al. (2015) raised questions for future research, especially regarding the possibility that the relation between right-wing authoritarianism and support for the prohibition of hate speech might be more gradual. For instance, when authoritarians face high negative emotional arousal or psychological threat and, as a result, can no longer control their behavior or adhere consistently to cultural norms, they might present less support for the hate speech prohibition.

This hypothesis is proposed because Bilewicz et al. (2015) identified differences in authoritarians' responses to hate speech directed against minorities whom they consider threatening versus the ones whom they consider harmless. However, this issue has not been addressed directly in their research.

1.3. The Influence of Social Norm on Hate Speech

Previous studies found a relationship between perceived social norms and prejudice expression in offline settings (Crandall et al., 2002; Crandall & Eshleman, 2003). In addition, the tendency of strong adherence to social norms might influence students not to accept prejudice toward minority groups but allow prejudice towards racism. In other words, conformity to an anti-discriminatory social norm accepts prejudice towards the deviant group (discriminatory) (Crandall et. al. 2002).

A similar study investigated the impact of perceived social acceptability on online hate speech. In addition, researchers measured the causal effect of counter-speaking and deleting hateful content (Álvarez-Benjumea & Winter, 2018). Results show that individuals infer acceptability from the context, using previous actions as a source of normative information. Simply put, when people observe that others have violated a certain social norm, such as using hate speech, they are likely to engage in similar behavior, in this case, to make use of hostile speech as well. Thus, social norms may create a fear of being excluded from a group, and to avoid being ostracized an individual may mimic others by using or not using hate speech.

This study by Álvarez-Benjumea & Winter (2018) considers that people learn about norms by observing others or by observing norm violations being sanctioned (censoring hate speech comments)

The experimental design comprised an online forum where German residents participated (n=180) could discuss current social topics in a manipulated environment. Participants were asked to join the conversations and leave comments. The experiment manipulated the comments participants could see before writing their own comments. Four conditions were presented: 1) control, called Baseline by the authors (6 comments; 2 friendly; 2 neutral and 2 hostile); 2) censored (4 comments; 2 friendly; 2 neutral); 3) extremely censored (3 comments; all friendly); 4) counter-speaking (6 comments; 1 friendly; 1 neutral; 2 hostile and 2 sanctions). Whereas participants in the baseline condition saw a balanced mix of friendly, neutral, and hostile comments, in the censored condition the researchers deleted prior hate content and presented participants only with friendly and neutral comments. In the extremely censored condition, they presented only friendly comments. Information on whether comments had been deleted was not displayed. In the counter-speaking condition, the hostile comments were presented with replies highlighting the unacceptability of hostile opinions, for example, "this is a prejudiced judgment" (injunctive norm).

The researchers collected the participants' own comments (n=1.555) and found that people tend to follow social acceptability norms they inferred from censored moderation of previous comments, even when others are unknown, and the speaker remains anonymous on the platform. "Participants were less likely to make use of hostile speech when they were presented with an environment in which previous extreme hate content had been censored" (Álvarez-Benjumea & Winter, 2018, p. 11). According to Álvarez-Benjumea & Winter's (2018) finding, descriptive norm (censorship condition) is more effective than injunctive norms (counter-speaking condition):

> Our findings contribute to the sociological literature in social norms by raising the question of whether descriptive norms might, in some settings, be more effective than sanctions at preventing antisocial behavior. Our results suggest that normative behavior in online conversations might, in, fact, be motivated by descriptive norms rather than injunctive norms (p. 11)

These results do not fit with the preexisting theory of social norms that defends those sanctions on normative behavior as more effective (Fehr & Gachter, 2000).

Bilewicz & Soral (2020) revealed that being previously exposed to hate speech has effects on the emotional, behavioral, and normative levels, creating a "desensitization to hate speech"

(Bilewicz & Soral, 2020, p.6). In this study, the authors explain mechanisms that have the effect of inhibiting the spread of Hate Speech, such as authoritarianism, social norm, and empathy.

This literature suggests that individuals with strict law and order rules are more likely to oppose and ban hate speech on social networks, even if they have high levels of prejudice that would go against this behavior (Bilewicz et al., 2015). Authoritarian individuals tend to emphasize law and order and promote normative conduct. If authoritarians consider hate speech to be deviant from social norms, they will be more likely to support the prohibition of hate speech. Social norms may moderate these results as previous studies have shown (Álvarez-Benjumea & Winter, 2018; Bilewicz & Soral, 2020).

1.4. Authoritarianism Might Prevent Hate Speech?

The theory of the Authoritarian Personality, developed in the mid-part of the 20th century by Adorno et al. (1950), suggests a set of personality traits that characterize authoritarian personalities, and the author developed the so-called ‘F-scale’ to measure their intensity. In the 1980s, Adorno’s theory served as an inspiration for new researchers, most notably Altemeyer´s (1998) and his concept of Right-Wing Authoritarianism (RWA).

The research identified that some individuals are more susceptible to authority figures, conform to societal conventions and norms, and are punitive towards minority groups that do not follow those norms. Accordingly, Altemeyer (1998) identified three dimensions of RWA: authoritarian submission, conventionalism, and authoritarian aggression.

In the past years, a substantial body of research has arisen using the RWA scale or variations of it. (Bizumic & Duckitt, 2018; Duckitt, 2020; Duckitt et al., 2010; Duckitt & Sibley, 2016). Research has shown that individuals with high scores on the authoritarianism scale are also high in prejudice and more socially conservative and nationalistic and that those individuals prefer strict rules and social control. People with a lower score on RWA are generally more tolerant and liberal, favoring individual liberties, high levels of personal freedom, self-expression, individual self-regulation, and support for democracy (Bizumic & Duckitt, 2018; Duckitt, 2020; Duckitt et al., 2010; Duckitt & Sibley, 2016).

According to Bilewicz et al.’s study (2015), the positive correlation between hate speech prohibition and right-wing authoritarianism is particularly strong with those groups that are protected by political correctness norms. In the Poland context, where the study was conducted, more protected groups are Africans and Ukrainians. However, the researchers noticed that the positive relation between right-wing authoritarianism and hate speech prohibition was not that strong with groups that are less protected in the Polish society, like LGBT and Muslims. These results support the perspective that authoritarians are willing to confront hate speech as long as those statements are against the laws and socially established rules.

Regardless of these findings, previous research has consistently found positive correlations between right-wing authoritarianism and prejudice. For instance, Sibley & Duckitt (2008) concluded in their meta-analysis of 71 studies (N = 22068) on the relationship between personality traits, RWA, Social Dominance Orientation (SDO), and prejudice that their "…findings confirmed the well-established conclusion in the research literature that RWA and SDO are strong predictors of prejudice, with the effect of each substantially independent of the other" (Duckitt & Sibley, 2010, p. 19). According to Duckitt & Sibley (2007), right-wing authoritarianism may predict prejudice toward groups that are seen as threatening the in-group's values, norms, and security and as potentially disrupting social stability and cohesion (Duckitt & Sibley, 2007). So right-wing authoritarianism may predict prejudice, especially against groups considered as socially deviant, people that are seen as threatening to the established norms and values of society. This relation between right-wing authoritarianism and prejudice is less strong regarding outgroups that are seen as socially subordinate. Nevertheless, the robust RWA-prejudice link suggests that individuals with a high score on right-wing authoritarianism should support hate speech against minority groups.

To sum up, previous research suggests that there should be an ambivalent relation between authoritarianism and hate-speech prohibition. On the one hand, authoritarianism should increase support for hate speech due to its relation to prejudice. On the other hand, Bilewicz et al.'s (2015) research confirms that there is an understanding of hate speech as a violation of a social norm. As a result, people with a high score on the RWA scale, despite carrying prejudiced attitudes, are also averse to hate speech, and as an outcome support hate speech prohibition.

For the current study, we focus on the RWA scale and its relation with the support of hate speech prohibition (Bilewicz et al., 2015) and examine whether this relation depends on social norms. Additionally, we investigate the interaction between right-wing authoritarianism and threat, that is, analyze if there are differences in support of the prohibition of hate speech against groups seen as more threatening and, groups seen as less threatening.

1.5. Interaction Between Perceived Threat and Right-Wing Authoritarianism

Inspired by Bilewicz et al.'s (2015) proposal for future research, we investigate whether the psychological threat has an impact on the relation between authoritarianism and hate speech prohibition. Previous studies have found that threat has a strong correlation with right-wing authoritarianism (Cohrs & Asbrock, 2009; Doty et al., 1991; Onraet et al., 2015; Vallejo-mart & Canto, 2021; Willis-Esqueda et al., 2017). As presented in the previous paragraphs, according to the literature review, right-wing authoritarianism should drive prejudice, especially against groups threatening the social order, stability, and safety.

The study by Cohrs & Asbrock (2009), with German students (n=176), found that when the outgroup was manipulated to appear socially threatening, right-wing authoritarianism had a powerful effect on prejudice. The experiment used three experimental conditions: threat, competition, and control. In the Competition condition, the Turkish (outgroup) employee candidates were presented with great power and status to compete with Germans in the job market. In the Threat condition, participants were informed about the growing number of Turks and Turkish students in Germany, the increasing influence of Islam in everyday life, criminality in Turkey, and rising crime rates in the country. In the control condition, the interviewee did not refer to social threat or competitiveness. Measures were both the SDO scale and the RWA scale, as well as the German version of the subtle and blatant Prejudice scale.

Right-wing authoritarianism was a strong overall predictor of prejudice and interacted marginally with experimental conditions for both blatant and subtle prejudice. “In both studies, right-wing authoritarianism was a powerful predictor of prejudice, but particularly so with regard to a threatening outgroup, compared with a group competitive for power-status or a control group” (Cohrs & Asbrock, 2009, p.284).

Considering the interaction between right-wing authoritarianism and threat, we manipulate threat by the target group in our research and analyze whether the relation between right-wing authoritarianism and support of hate speech prohibition will be reduced in the threat condition as compared to a control condition. We expected that that should be the case because the positive relation with hate speech prohibition, which is due to high RWA individuals’ motivation to maintain the social law and order, will be counteracted by a negative relation with hate speech prohibition due to increased prejudice to the threatening target group. Thus, the threat is going to be a moderator.

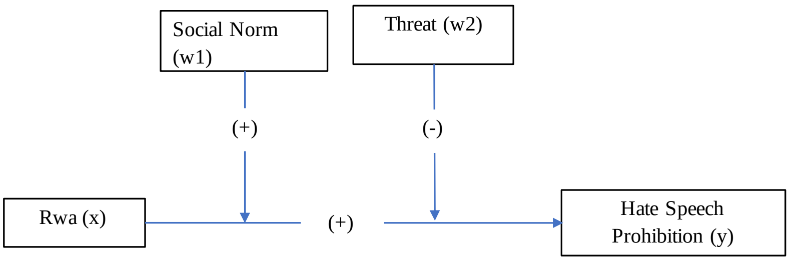
1.6. Present Study

The current study examines whether ideological beliefs affect the support of hate speech prohibition. More precisely, replicating previous findings, the current study tests if authoritarianism predicts confrontation against hate speech on social media. The study adds to the existing research by examining whether threat and social norms moderate the relation between right-wing authoritarianism and hate speech prohibition. Considering previous findings regarding group hate speech, authoritarianism, perceived threat, and social norms, we formulate the following theoretical model and hypotheses (Figure 1.1).

1.1.1. Theoretical Model

Figure 1.1

Theoretical Model Framework



Note. RWA as predictors of hate speech prohibition and social norms and threat as moderators. The sign "-" indicates a negative relationship and the sign "+" indicates a positive relationship.

1.1.2. Hypotheses

Based on the theoretical reasoning outlined above and taking into existing evidence for the moderating role of social norms and threat, we propose the following hypotheses:

H1: Right-wing authoritarianism (X) positively predicts support for the prohibition of hate speech (Y).

H2: The normativity of hate speech moderates the relation between right-wing authoritarianism (X) and hate speech prohibition, with a stronger positive relation between right-wing authoritarianism and support of hate speech prohibition (Y) when hate speech is non-normative as compared to when it is normative.

H3: Threat moderates the relation between right-wing authoritarianism (X) and hate speech prohibition (Y), with a more positive relation under low threat than under high threat.

To test our hypotheses, we run an online experiment in which we presented 16 examples of hate speech and asked participants whether they would permit, prohibit, or report these hate speech expressions. To study hate speech against target groups in Portugal, we used two known target groups, the LGBT community, and Immigrants. This selection was made after a pre-interview with Facebook and Instagram content moderators, who reported to us which groups were the most targeted. We measured RWA and manipulated both, threat by the group targeted and the normativity of hate speech

CHAPTER 2

Method

2.1. Method

2.1.1 Design and Procedure

The study had a 3x2 mixed factor design, with the participants being randomly assigned to one of three conditions of the threat manipulation (control condition vs. LGBT threat vs. immigrants' threat) as a between-subjects factor. Each of the hate-speech examples that were presented to each participant was randomly assigned to one of the conditions of the manipulation of the normativity of hate speech (Censored x Non-censored) as a within-subject factor. The dependent variable was the hate speech prohibition. Right-wing authoritarianism was measured as an independent variable.

The study consisted of four parts:

First, we assessed participants' nationalities and if they have been living in Portugal for the last two years. Second, depending on the experimental condition the participant was asked to read two newspaper articles about one of the two target groups (LGBT threat, immigrant threat) or sport and a farmer's manifestation (control condition). The newspapers were presented in a Twitter post layout, to simulate an experience of using social networks (See Appendix B, 7.1, 7.2, and 7.3).

After answering some filler questions ("Did you read about this news before?" and "Have you seen this news on social media"?) about the newspaper posts, participants responded to the RWA scale, two Social Distance scales (one for each target group), and the LGBT and immigrant threat scale.

Next, participants were asked to read examples of hate speech and to indicate for each example to which degree they agreed with prohibiting, permitting, or reporting the presented online hate speech. In this part, participants received additional information about the hate speech, depending on the condition of the normativity manipulation (censored vs. non-censored). Finally, participants' sociodemographic information was assessed and their experience of illegal content online and their opinions about the role and responsibility of online platforms were measured with the Media Monitors scale

2.2. Manipulation

2.2.1. Threat Manipulation

The manipulation of threat consisted of asking participants to read carefully two (2) Twitter posts with news, from Portuguese newspapers, one from *Público* and the other one from *Observador*. The layout of the news was taken from a Twitter post. (See Appendix B, Figures 2, 3, and 4).

In the Immigrant Threat condition, the first news post was related to high percentages of cases of the sexually transmitted disease HIV in the immigrant population in Portugal. The second news was about the conquest of rights, elaborating on the fact that the children of immigrants living in Portugal can already have their Portuguese nationality.

In the LGBT threat condition, the first news post was also related to HIV and World Health Organization (WHO) recommendations for gay men. The second news post was about the conquest of rights, elaborating on the fact that Portugal had two years before legalized the adoption of children by homosexual couples.

In the control condition, one news post was about soccer, and the other was about a manifestation of farmers.

2.2.2. Normativity Manipulation

To manipulate the normativity of hate speech, two groups of hate speech examples were presented: The first group (censored condition) consisted of hate speech examples that were presented with a warning about an infraction of community guidelines: “your publication was removed” with the additional information: “we removed your post because it does not comply with our Community Guidelines. If you violate our guidelines again, your account may be restricted or disabled[2]”; The second group (non-censored) consisted of hate-speech examples that were presented without such added statements (See Appendix B, Figure 1).

All participants in this study were present with censored and non-censored hate speech examples (within-subject manipulation of normativity). The prediction was that the positive relation between right-wing authoritarianism and hate speech prohibition will be stronger in the censored condition (low normativity) than in the non-censored condition (control).

In total, each participant saw 16 Hate Speech examples divided into two separate blocks, one with eight examples of hate speech against immigrants and the other with eight examples of hate speech against LGBT. Whether each hate-speech example was presented as censored or non-censored was randomized for each participant, as was the order of target group and the

[2] Portuguese translation: “Sua publicação foi excluída. Removemos a sua publicação porque ela não cumpre as nossas Community Guidelines. Se você infringir as nossas diretrizes novamente, a sua conta poderá ser restrita ou desativada”.

order of hate speech examples for each target group. Thus, each participant saw all the 16 hate-speech examples, each one either censored or non-censored, but never the same item twice (censored and non-censored)

2.2.3. Selection of Hate Speech Expressions

To proceed with the study, we selected examples of hate speech based on interviews with four volunteers who work or have worked, in Portugal, as Content Moderators for Facebook, Instagram, and YouTube. From these interviews, we collected 21 real online hate speech statements.

From this database, the most representative statements were directed against stigmatized groups in Portugal: Immigrants (in most cases from countries like Brazil, China, African countries with Portuguese as the official language: Países Africanos de Língua Oficial Portuguesa – PALOP – and Indians, Black People, Lesbian, Gay, Bisexual, and Transgender (LGBT Community), Women, and Romas.

For the current study, we used hate speech examples of two categories: Hate speech against immigrants and hate speech against LGBT. Also, the criteria for selecting those expressions were based on UNESCO guidelines, considering that the UN Strategy and Plan of Action of hate speech refers to the definition as: “any kind of communication in speech, writing or behavior that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor” (UNESCO, 2022, June 16).

Our definition of this phenomenon, which is still poorly understood, included recommendations from the UNESCO report *Countering Online Hate Speech* (Gagliardone et al., 2015). According to Gagliardone et al. (2015), narrow definitions of hate speech fail to consider the dangers of amplifying violence, especially in online hate speech, as well as the potential to cause harm and violence. In addition to calling that character "dangerous speech", the author also commented on "fear speech", which emphasizes the fear of a group. In the authors words about online hate speech “…all definitions still incur the intractable challenge of making connections between the online expressions of hatred and actual harm such as hostility, discrimination, or violence” (Gagliardone et al., 2015, p. 54).

In total, we selected 16 statements and adapted those statements for eight for LGBT and eight for Immigrants (See Appendix A, Table 7).

2.2.4. Participants

Participants were recruited via e-mail and social media platforms, using snowball sampling. A link to an online survey on the Qualtrics platform was sent to individuals. An informed consent stating that the study is voluntary, anonymous, and confidential was displayed to all participants on the first screen of the survey. The survey was conducted in Portugal. At the end of the questionnaire, participants could take part in a raffle of two gift cards from Celeiro and a debriefing page was shown to the participants that emphasized the artificial character of the presented material and explained in more detail the objectives of the research. It also provided contact information in case participants wanted to inquire more about the study.

A total of 291 people responded to the questionnaire, 86 participants did not meet the inclusion criteria and were excluded from the analysis. The criteria were to be Portuguese or be living in Portugal for the last two years and to have completed the full survey. Of the 205 remaining participants, 167 (81.5%) were Portuguese, 37 (18%) were from other nationalities and one participant did not reveal the nationality but had been living for the last two years in Portugal.

The range of participants' age was 18-64 years with most participants 51.2% aged between 25-34 years (n = 105) (See Table 2.1).

Table 2.1

Q. How Old Are You?

		Frequency	Percent
Valid	18 - 24	14	6.8%
	25 - 34	105	51.2%
	35 - 44	57	27.8%
	45 - 54	18	8.8%
	55 - 64	4	2.0%
	Total	198	96.6%
Missing[a]	-99	2	1.0%
	System	5	2.4%
	Total	7	3.4%
Total		205	100.0

Note. [a] = participant that did not answer the question *How old are you?*

Seventy-seven participants (37.6%) were female, and 122 participants (59.5%) were male, one participant was non-binary, and five participants did not indicate their gender (2.40%). In our sample, 119 participants (58%) have a university degree (Bachelor, Master, or

PhD.), 49 (23.9%) participants did not conclude the university yet, two (1%) did less than junior high school, 28 (13.7%) finished high school and 14 (6.8%) participants did not answer.

Most of the participants 157 (76.6%), were heterosexual, 20 (9.8%) were homosexual and 20 (9.8%) were bisexual, one participant chooses the “others” option of this question, and two participants indicated that they preferred not to answer, and five (2.40%) participants did not answer.

Political views ranged from 1 to 7 on a 7-point scale when 1 is *left-wing* and 7 is *right-wing*. Forty-seven (22.9%) participants marked from 3 to 1 on the scale, identifying themselves more as left-wing, and 87 (42.4%) participants marked from 5 to 7 on the scale, identifying themselves more as right-wing. Fifty-one (24.9%) participants marked 4 on the scale and 20 (9.8%) participants did not answer.

Regarding the frequency of social media use, 97 participants (47.3%) use social media *daily*, 20 (9.8%) participants use social media from *4 to 6 times a week*, 14 (6.8%) use social media from *2 to 3 times a week,* 5 (2.40%) use social media *1 time a week*, 1 (1.0%) *never used social media*, and 68 participants (30.7%) did not answer

2.3. Measures

If not described otherwise, scales of all measures were presented as visual analog (VAS) scales with answer options ranging from 0 (*completely disagree*) to 100 *(completely agree*). The benefit of using VAS is to allow respondents to freely specify the position of their perceived status instead of being limited to certain predetermined categories (Chang & Little, 2018).

2.3.1. Right-Wing-Authoritarianism

Right-Wing Authoritarianism (RWA) was measured on a short scale from Duckitt & Bizumic (2018). The Very Short Authoritarism (VSA) scale was adapted to Portuguese (See Appendix A, Table 1). Therefore, it was translated to Portuguese by the author and translated back into English by two other bilingual volunteers. This back-translation was then compared to the English original version.

The instrument was a set of six items. Items one, four, and five were reverse coded.

To test the validity of the Portuguese translation, a factor analysis of the six items was conducted with Maximum Likelihood extraction and Direct Oblimin rotation. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = .646. Bartlett’s test of sphericity, x2 (15) = 519, p < .001, indicating that the correlation structure is adequate for factor analyses. The maximum likelihood factor analysis with a cut-off point of .40 and the

Kaiser's criterion of eigenvalues greater than 1 (Hair et al., 2010) yielded a two-factor solution as the best fit for the data, accounting for 73.2% of the variance, seen in Table 2.2. Consistent with that, the curve in the scree plot also showed 2 factors before the more substantial drop in explained variance, as seen in Figure 2.1. Both factors were uncorrelated (r = .037). The results of this factor analysis are presented in Table 2.3.

Table 2.2

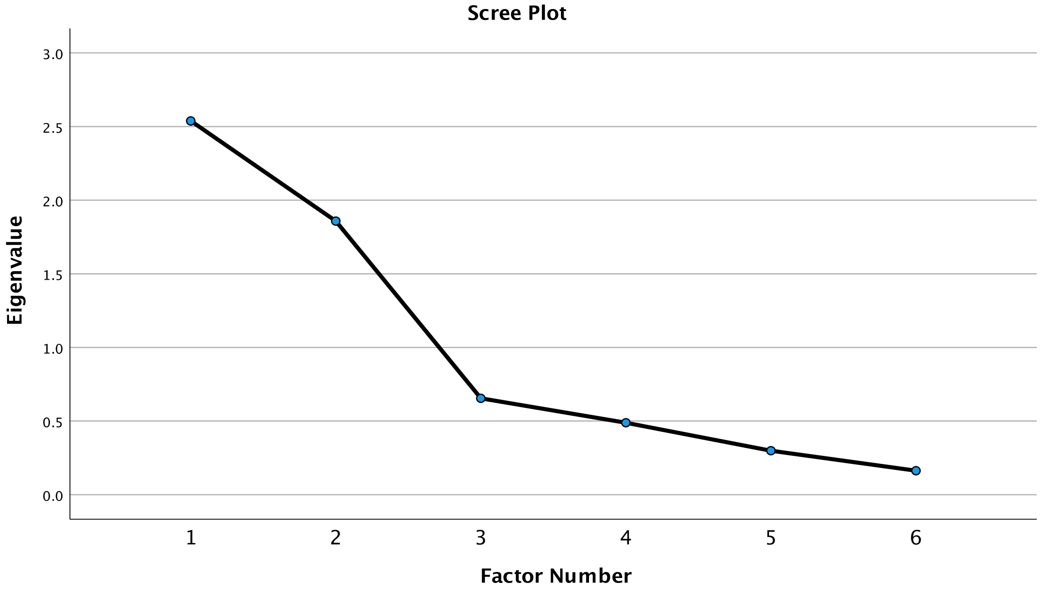
Total Variance Explained Table of the Factor Analysis of the Very Short Authoritarianism Scale

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	**2.53**	**42.3**	**42.3**	2.25	37.5	37.5	2.26
2	**1.85**	**30.9**	**73.2**	1.35	22.5	60.1	1.34
3	.654	10.9	84.1				
4	.489	8.14	92.3				
5	.299	4.97	97.2				
6	.163	2.70	100				

*Note. The e*xtraction method was maximum likelihood with an oblimin (Promax with Kaiser normalization) rotation. a. When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

Figure 2.1

Scree Plot from RWA Factor Analysis



Nevertheless, as the original scale had been validated before and because the distinction of the two factors was only because half of the items were reversed coded, we decided to create the overall RWA score by averaging the ratings across all items (Cronbach's alpha = .61).

Table 2.3

Pattern Matrix Table of the Factorial Analysis of the Very Short Authoritarianism Scale

	Factor	
	1	2
O que nosso país mais necessita é de disciplina, com todos a seguir os nossos líderes em unidade	**.969**	-.106
A lei de Deus sobre aborto, pornografia e casamento tem de ser estritamente seguida antes que seja tarde demais.	**.803**	.127
Os fatos sobre o crime e as recentes desordens públicas mostram que temos que reprimir mais duramente os criadores de problemas, se quisermos preservar a lei e a ordem.	**.778**	-.044
A nossa sociedade NÃO precisa de um governo mais rigoroso e de leis mais rígidas. (R)	-.040	**.720**
É ótimo que muitos jovens hoje estejam preparados para desafiar a autoridade. (R)	.242	**.675**
NÃO há nada de errado com a relação sexual antes do casamento. (R)	-.129	**.580**

*Note. N = 205. The e*xtraction method was maximum likelihood with an oblimin (Promax with Kaiser normalization) rotation. Factor loading above .30 are in bold. Reverse-scored items are denoted with (R). a. Rotation converged in 3 iterations.

Although it would have been possible to slightly increase the internal consistency of the scale by removing one item from Cronbach's alpha =.614 to .685, we decided to keep all items, because the scale is already very short (See Appendix A, Table 4).

2.3.2. Outgroup Prejudice

For each of the two target groups (immigrants and LGBT) outgroup prejudice was measured with a Social Distance scale (Mather et al., 2017). Participants had to answer whether they would accept a member of the target group as a coworker, as a neighbor, or as a marriage partner of a relative. A Portuguese validated scale from the European Value Survey – EVS, (2008) was used. Originally, the Portuguese Social Distance scale included three items, with immigrants as target groups. The same items were used to measure Social Distance from LGBT by changing "immigrants" to "homosexuals" (i.e., item 1: "Em que medida se sentiria incomodado se um homossexual fosse seu vizinho de rua"; item 2: "Em que medida se sentiria incomodado se um homossexual fosse nomeado seu chefe" and item 3: "Em que medida se sentiria incomodado se um homossexual casasse com um familiar próximo"). The overall Cronbach's alphas were .90 for immigrants and .95 for LGBT, showing high levels of consistency (See Appendix A, Table 1).

The mean of the three items for each of the target groups was computed for the final scores of prejudices against immigrants and against the LGBT community (r= .869, $p < 0.001$) We also computed an overall prejudice score as the average of prejudice against immigrants and against the LGBT community

2.3.3. Threat

Perceived threat by immigrants was measured with the Intergroup Threat scale (Stephan et al., 1999). A Portuguese validated scale from the European Social Survey (ESS7) was used with four items. For measuring LGBT threat, we adapted four items of the Perceived Threat of Homosexuals scale (Tjipto et al., 2019).

The items were translated to Portuguese by the author and translated back into English by two other bilingual volunteers, (i.e., item one: "Os homossexuais prejudicam os costumes, as tradições e a vida cultural em Portugal"; item two: "Os homossexuais possuem valores e crenças que representam uma ameaça às questões morais e religiosas em nossa sociedade"; item three: "Os homossexuais contribuem para a diminuição da população"; and item four: "Os homossexuais aumentam os níveis de doenças sexualmente transmissíveis em nossa sociedade portuguesa" (See Appendix A, Table 2).

To test the validity of the translation a factor analysis of the 4 items was conducted with Maximum likelihood extraction and Direct Oblimin rotation. A one-factor solution showed a good fit for the scale, with factor one responsible for 81.5% of the variance (See Appendix A, Table 3).

The overall Cronbach's alphas were .95 for threat by immigrants and .96 for threat by LGBT, showing high levels of consistency (See Appendix A, Table 1).

Threat scores for the two target groups were calculated as the mean of the 4 items of LGBT threat and, immigrant threat (r = .863, p < 0.001). An overall threat measure was computed as the average of LGBT threat and immigrant threat

2.3.4. Hate Speech Scale

To measure hate speech prohibition, each participant was asked for each of the 16 hate-speech examples if they consider the statement should be prohibited, permitted, or reported on the platform with answer options ranging from 0 (*completely disagree*) to 100 (*completely agree*). The layout of the Hate Speech statements was a Facebook post (see Appendix B, Figure 1).

For the calculation of Hate Speech Prohibition scales, we averaged responses to the hate-speech examples for each target group (immigrants and LGBT), each normativity condition (censored vs non-censored), and item (prohibiting, permitting, and reporting). Responses on the permitting items were reversed coded.

In the end, 12 variables were obtained, namely.

1) Hate Speech against LGBT:

a – HC1: Probihition of HateSpeech – censored;

b - HC2: Permission of HateSpeech – censored (reversed coded);

c – HC3: Report of HateSpeech – censored;

d - HS1: Probihition of HateSpeech – non-censored;

e – HS2: Permission of HateSpeech – non-censored (reversed coded);

f – HS3: Report of HateSpeech – non-censored).

2) Hate Speech against Immigrants:

a – IC1: Probihition of HateSpeech – censored;
b – IC2: Permission of HateSpeech – censored (reversed coded);
c – IC3: Report of HateSpeech – censored;
d – IS1: Probihition of HateSpeech – non-censored;
e – IS2: Permission of HateSpeech - non-censored (reversed coded);
f – IS3: Report of HateSpeech – non-censored.

Although we initially considered calculating one overall Hate-Speech-Prohibition Scale based on the average of ratings on all three items, Cronbach's alphas for such a scale using "prohibition", "permission" and "report" measures were not acceptable. For that reason, we treated responses to the three items separately in the data analyses (See Table 2.4). There was a strong correlation between immigrant hate speech prohibition censored and non-censored, *r* (205) = .933, p <.001, and a strong correlation between LGBT hate speech prohibition censored and non-censored, *r* (205) = .943, p <.001. (See Pearson correlations in Appendix A, Table 6).

Table 2.4

Descriptive Repeated Measures Labels on Target Group and Normativity Within-Subjects Factors

Target	Normativity	item	Dependent Variable
Immigrant	Censored	1	IC1
		2	IC2_r
		3	IC3
	Non-Censored	1	IS1
		2	IS2_r
		3	IS3
LGBT	Censored	1	HC1
		2	HC2_r
		3	HC3
	Non-Censored	1	HS1
		2	HS2_r
		3	HS3

Note. IC1= Immigrant Censored Hate Speech Prohibition; IC2_r = Immigrant Censored Hate Speech Non-Permission; IC3 = Immigrant Censored Hate Speech Report; IS1 = Immigrant Non-Censored Hate Speech Prohibition; IS2_r = Immigrant Non-Censored Hate Speech Non-Permission; IS3 = Immigrant Non-Censored Hate Speech Report; HC1 = LGBT Censored Hate Speech Prohibition; HC2_r = LGBT Censored Hate Speech Non-Permission; HC3 = LGBT Censored Hate Speech Report; HS1 = LGBT Non-Censored Hate Speech Prohibition; HS2_r = LGBT Non-Censored Hate Speech Non-Permission; HS3 = LGBT Non-Censored Hate Speech Report

2.3.5. Sociodemographic Data and Media Monitoring

Regarding demographics, gender, age, education, nationality, and sexual orientation were assessed. In addition, political orientation was measured ranging from 1 (*left-wing)* to 7 (*right-wing*). Moreover, we used the Media Monitoring Scale (European Commission, 2018) to explore respondents' experience of illegal content online and their opinions about the role and responsibility of online platforms. The items were: 1) *A Internet é segura para os seus utilizadores* ("The Internet is safe for its users"); 2) *É necessário tomar medidas para limitar a disseminação de conteúdo ilegal na Internet* ("It is necessary to take measures to limit the dissemination of illegal content on the Internet"); 3) *A liberdade de expressão precisa de ser protegida online* ("Freedom of expression needs to be protected online"); 4) *Os serviços de alojamento na Internet são eficazes a lidar com conteúdo illegal* ("Internet hosting services are the solution to dealing with illegal content").

CHAPTER 3

Results

3. 1. Results

SPSS v28.0 statistical software was used to perform the data analysis in this study. First, the reliability of each of the scales used was calculated. Then, the descriptive statistics (average and standard deviation) and the correlation between variables were analyzed. Finally, GLMs with repeated measures were carried out.

The descriptive statistics (mean and standard deviation) are presented in Appendix A, Table 5, and the correlation between variables is presented in Appendix A, Table 6. Not surprisingly, the results reveal a positive correlation between right-wing authoritarianism (RWA) and outgroup prejudice, *r* (205) = .429, *p* <.001, and a strong correlation between right-wing authoritarianism and perceived threat, *r* (205) = .492, *p* <.001.

We expected that the influence of authoritarianism on hate-speech prohibition should be less positive in the case of groups that are being perceived as threatening. To test this hypothesis, we had manipulated threat by creating three treatment conditions: 1) immigrant threat; 2) LGBT threat, and 3) control condition. Table 3.1 shows the descriptive statistics of measured threat by the two target groups in the different experimental conditions.

Table 3.1

Perceived threat by LGBT and by Immigrants in the 3 conditions of the threat manipulation (Immigrant vs LGBT vs Control).

	Condition	Mean	Std. Deviation	N
Perceived Threat by LGBT	Immigrant	41.8	28.1	75
	LGBT	38.3	28.8	62
	Control	37.0	27.8	68
	Total	39.2	28.2	205
Perceived Threat by Immigrants	Immigrant	38.7	28.0	75
	LGBT	36.2	29.3	62
	Control	35.2	28.4	68
	Total	36.8	28.4	205

3.1.1. Manipulation Check

To test whether the manipulation of threat affected the perceived threat of those target groups, data were analyzed using a 2 (threat target: LGBT vs. Immigrants) x 3 (threat condition: Control vs. LGBT vs. Immigrants) mixed-design GLM with threat target as within-subjects factor and threat condition as between subjects' factor on the threat measures as dependent variables. Mauchly's test indicated that the assumption of sphericity had been violated ($\chi 2(0)$, $p < .001$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 1.00$). The main effect of threat target was significant, $F(2, 202) = 5.08$, $p = .025$, $eta_p^2 = .025$, indicating that participants felt more threatened by the LGBT community ($M = 39.22$, $SD = 28.21$) than by immigrants ($M = 36.83$, $SD = 28.48$). This main effect was not qualified by the predicted interaction with condition, $F(2, 202) = .165$, $p = .848$, $eta_p^2 = .002$, indicating that the threat-manipulation was probably not successful.

We cannot conclude that we had a success in the manipulation of threat. The main effect of condition was also not significant, $F(2, 202) = .440$, $p = .644$, $eta_p^2 = .004$

3.1.2. Effects on Hate Speech Prohibition

To test whether right-wing authoritarianism would positively predict support of hate speech prohibition (H1) and whether normativity of hate speech (H2) and threat (H3) moderates the relation between right-wing authoritarianism and hate speech prohibition (H2), we run a 3 (threat condition: Control vs. LGBT threat vs Immigrant threat) x 2 (target: Immigrants vs. LGBT) x 2 (normativity: Censored vs. Non-censored) x 3 (items: Prohibition vs. Non-permission vs. Reporting) mixed GLM with the condition as between-subjects factor and all others as within-subjects factors, using the immigrant hate speech prohibition items (IC1, IC2_r, IC3, IS1, IS2_r, IS3) and LGB hate speech prohibition items (HC1, HC2_r, HC3, HS1, HS2_r, HS3) as dependent variables and mean-centered right-wing authoritarianism as a continuous predictor.

The interaction between threat condition and centered RWA as well as its higher order interactions with the within-subjects factors were included in the model. Mauchly's test indicated that the assumption of sphericity had been violated ($\chi 2(2) = .669$, $p < .001$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .751$).

3.1.3. RWA and Hate Speech Prohibition (H1)

The main effect of right-wing authoritarianism was significant, $F(1, 199) = 60.2$, $p < .001$, $\eta p2 = .232$. However, parameter estimates indicated that it was negative, contrary to what was predicted (Table 3.3). That is, the higher participants scored on RWA, the less they supported hate-speech prohibition, and this was the case for all three types of items (prohibiting, not-permitting and reporting).

3.1.4. Moderation by Normativity (H2)

The main effect of normativity was not significant (Table 3.2). However, the unpredicted negative effect of RWA was qualified by a weak but significant interaction between normativity and right-wing authoritarianism *F(*1,199) = 5.71, *p* = .018, ηp2 = .028. For LGBT targets this interaction was partially consistent with H2, because, although there was no positive effect of RWA, at least the unpredicted negative relation between RWA and hate speech prohibition was weaker in the censored than in the non-censored condition. The picture was less clear for immigrants, with different results on the different items (Table 3.3). However, the normativity by RWA interaction did not interact with target, nor with any other of the remaining factors or their combinations (Table 3.2)

3.1.5. Moderation by Threat Condition (H3)

Condition had no main-effect, *F*(2, 199) = 0.51, *p* = .60, η_p2 = .005, and did not interact with RWA, *F*(2, 199) = 2.02, *p* = .135, η_p2 = .020. More importantly, based on H3 we predicted a three-way interaction between RWA, experimental condition, and target because RWA should show a more positive relation with hate-speech prohibition in the condition in which the respective target group is less threatening. This interaction was not significant (Table 3.2). Thus, there is no evidence for H3. However, this result needs to be interpreted with caution, given that the manipulation check did not indicate a successful manipulation.

3.1.6. Unpredicted Effects

The main effect of items (prohibition vs. non-permission vs. reporting) and the interaction between items and right-wing authoritarianism were significant (Table 3.2). As can be seen in Table 3.3, for both target groups RWA had the strongest negative relation with the reporting of hate speech, the weakest negative relation with the prohibition of hate speech. The negative relation with non-permission was in between these extremes.

Additionally, the theoretically irrelevant (for the current research question) three-way interaction between items, normativity and condition was significant *F*(3.97, 395) = 3.16, *p* = .014, ηp2 =.031. All other main effects and interactions were non-significant and irrelevant to our hypotheses. See Table 3.2.

Table 3.2

Tests of Within-Subjects Effects of a GLM with hate speech prohibition items as dependent variable, threat-condition as between-subjects factor, mean-centered RWA as continuous predictor and target (LGBT vs. Immigrants), normativity (censored vs. non-censored), and item (prohibition vs. non-permission vs. reporting) as within-subject factors.

	df	*F*	*p*	η_p^2
target	1.000	0.006	.938	.000
target * Condition	2.000	1.938	.147	.019
target * RWA	1.000	0.016	.899	.000

target * Condition * RWA	2.000	0.610	.545	.006
Error(target)	199.000			
Normativity	1.000	0.601	.439	.003
Normativity * Condition	2.000	1.119	.329	.011
Normativity * RWA	**1.000**	**5.715**	**.018**	**.028**
Normativity * Condition * RWA	2.000	2.754	.066	.027
Error(Normativity)	199.000			
Item	**1.503**	**3.754**	**.036**	**.019**
Item * Condition	3.006	0.366	.778	.004
Item * RWA	**1.503**	**8.913**	**<.001**	**.043**
Item * Condition * RWA	3.006	0.252	.860	.003
Error(Item)	299.082			
target * Normativity	1.000	0.294	.588	.001
target * Normativity * Condition	2.000	0.379	.685	.004
target * Normativity * RWA	1.000	0.794	.374	.004
target * Normativity * Condition * RWA	2.000	0.598	.551	.006
Error(target*Normativity)	199.000			
target * Item	1.994	0.020	.980	.000
target * Item * Condition	3.989	1.024	.395	.010
target * Item * RWA	1.994	1.184	.307	.006
target * Item * Condition * RWA	3.989	0.366	.832	.004
Error(target*Item)	396.904			
Normativity * Item	1.987	0.092	.911	.000
Normativity * Item * Condition	**3.975**	**3.169**	**.014**	**.031**
Normativity * Item * RWA	1.987	1.758	.174	.009
Normativity * Item * Condition * RWA	3.975	0.369	.830	.004
Error(Normativity*Item)	395.473			
target * Normativity * Item	1.838	1.654	.195	.008
target * Normativity * Item * Condition	3.676	0.202	.926	.002
target * Normativity * Item * RWA	1.838	0.419	.641	.002
target * Normativity * Item * Condition * RWA	3.676	0.103	.976	.001
Error(target*Normativity*Item)	365.776			

Note. Greenhouse-Geisser correction applied to all effects
GLM = generalized linear model
* = interactions
Effects with *p*-values below .050 are in bold

Table 3.3

Parameter Estimates of the RWA main effects in a GLM with hate speech prohibition items as dependent variable, threat-condition as between-subjects factor, mean-centered RWA as continuous predictor and target (LGBT vs. Immigrants), normativity (censored vs. non-censored), and items (prohibition vs. non-permission vs. reporting) as within-subject factors.

						95% Confidence Interval		
Dependent Variable	Parameter	*B*	Std. Error	*t*	*p*	Lower Bound	Upper Bound	Partial Eta Squared
Immigrant								
IC1	RWA	-0.376	0.175	-2.14	.033	-0.722	-0.030	.023
IC2_r	**RWA**	**-0.663**	**0.185**	**-3.59**	**<.001**	**-1.02**	**-0.299**	**.061**
IC3	**RWA**	**-0.825**	**0.168**	**-4.91**	**<.001**	**-1.15**	**-0.494**	**.108**
IS1	RWA	-0.383	0.169	-2.25	.025	-0.717	-0.049	.025
IS2_r	RWA	-0.577	0.186	-3.10	.002	-0.944	-0.210	.046
IS3	**RWA**	**-0.807**	**0.166**	**-4.86**	**<.001**	**-1.13**	**-0.480**	**.106**
LGBT								
HC1	RWA	-0.341	0.178	-1.91	.057	-0.691	0.010	.018
HC2_r	RWA	-0.576	0.190	-3.02	.003	-0.951	-0.201	.044
HC3	**RWA**	**-0.729**	**0.175**	**-4.16**	**<.001**	**-1.07**	**-0.384**	**.080**
HS1	RWA	-0.376	0.172	-2.18	.030	-0.715	-0.037	.024
HS2_r	RWA	-0.584	0.186	-3.14	.002	-0.949	-0.218	.047
HS3	**RWA**	**-0.827**	**0.166**	**-4.98**	**<.001**	**-1.15**	**-0.500**	**.111**

Note. P values below .001 are in bold.
IC1= Immigrant Censored Hate Speech Prohibition; IC2_r = Immigrant Censored Hate Speech Non-Permission; IC3 = Immigrant Censored Hate Speech Report; IS1 = Immigrant Non-Censored Hate Speech Prohibition; IS2_r = Immigrant Non-Censored Hate Speech Non-Permission; IS3 = Immigrant Non-Censored Hate Speech Report; HC1 = LGBT Censored Hate Speech Prohibition; HC2_r = LGBT Censored Hate Speech Non-Permission; HC3 = LGBT Censored Hate Speech Report; HS1 = LGBT Non-Censored Hate Speech Prohibition; HS2_r = LGBT Non-Censored Hate Speech Non-Permission; HS3 = LGBT Non-Censored Hate Speech Report.

3.1.7. Additional Analyses

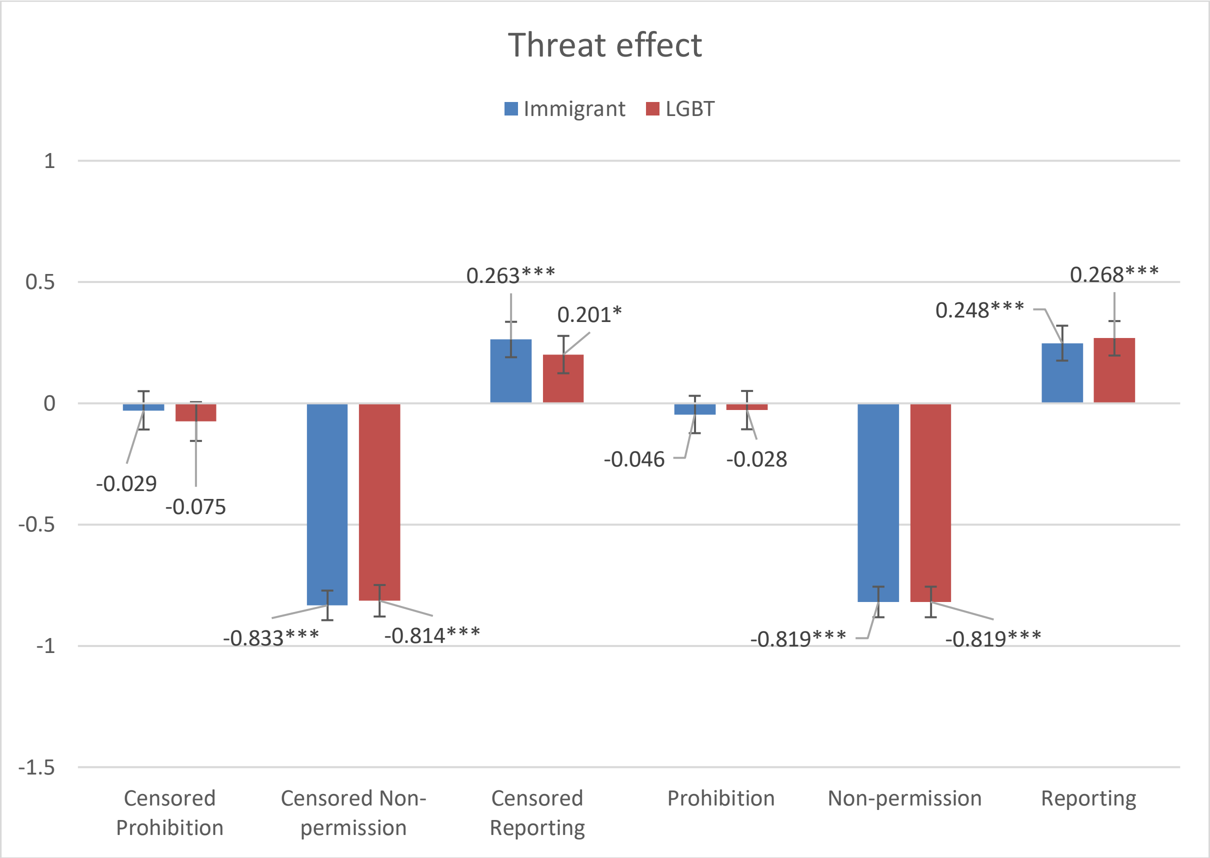
Given that the manipulation check had indicated that the manipulation was probably not successful, we tested hypothesis H3 by running the same analysis but with the measured perceived threat instead of manipulated threat as a moderator of the relation between RWA and hate-speech prohibition.

As the two measures of perceived threat by immigrants and by LGBT were very strongly correlated (r = .86) we created a threat-composite score by averaging the two. We then mean-centered this composite threat measure and included it as a continuous predictor and moderator in the GLM, substituting the threat conditions and keeping normativity, target, and items as within-subjects factors.

The strongest effect in this analysis was an unpredicted two-way interaction between perceived threat and items, *F*(1.719, 345.455) = 84.38, *p* < .001, η_p2 = .296, which was because perceived threat was negatively related to non-permission, positively related to reporting and unrelated to prohibition (see Figure 3.1). Thus, perceived threat increased the readiness to permit hate speech against both target groups, independent of normativity, but also the intention to report these hate-speech cases.

Figure 3.1

Effects of perceived threat on prohibition of censored and non-censored hate speech against LGBT and Immigrants



Note. Parameter estimates for threat effect shown for LGBT and Immigrant items in the two normativity conditions (censored, non-censored) in a 2 (normativity) x 2 (target group) repeated measures GLM on hate-speech prohibition, with mean-centered RWA, mean-centered perceived threat, and their interaction as continuous predictors. *p < 0.05; **p < 0.01; ***p < 0.001. Error bars represent standard errors.

RWA, *F*(1, 201) = 5.36, *p* = .022, η_p2 = .026, and perceived threat, *F*(1, 201) = 19.01, *p* < .001, η_p2 = .086, had significant main effects. More importantly, the interaction between RWA and perceived threat was also significant, *F*(1, 201) = 6.56, *p* = .011, η_p2 = .032. However, this two-way interaction was qualified by a three-way interaction with items, *F*(1.719,345.455) = 4.07, *p* = .018, η_p2 = .020, and a four-way-interaction of RWA, perceived threat, items and normativity, *F*(1.967,401.202) = 6.84, *p* = .001, η_p2 = .033.

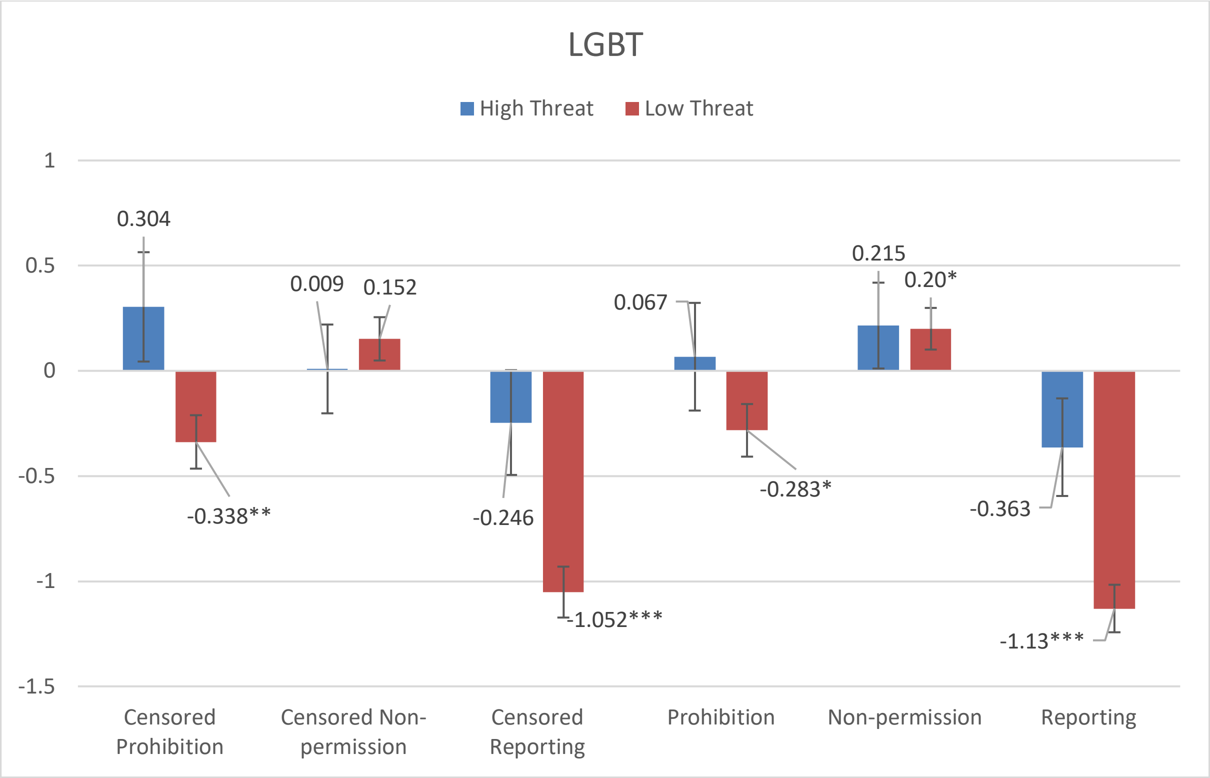
We then run the simple slopes-analysis to estimate the effects of RWA for participants high in perceived threat (one standard deviation above the mean) and those that were low in perceived threat (one standard deviation below the mean). More precisely, we run the same GLM but instead of the centered measure of perceived threat we included transformed threat measures with means equal to minus 1SD or plus 1SD in order to obtain parameter estimates for RWA effects under high versus low threat, respectively.

The parameter estimates analysis revealed that under high threat (1 SD above the mean; Table 3.4, upper part) the main effect of right-wing authoritarianism became positive and non-significant for items 1 (prohibition) and 2 (non-permission). For item 3 (reporting) the parameter of the RWA effect remained negative but was only significant in the censored condition and with immigrants as targets.

Different results were found for estimates at low levels of perceived threat (1 SD below the mean; Table 3.4 lower part). In this case, the main effect of RWA was again positive and mostly non-significant (only significant for non-censored condition in LGBT as target) for item 2 (non-permission) but remained negative and mostly significant for item 1 (prohibition), particularly in the censored condition. Moreover, the negative relation of RWA with item 3 (reporting) became much stronger than under high threat and was highly significant (See Table 3.4). For better visualization of the parameters, and estimates see Figures 3.2 and 3.3.

Figure 3.2

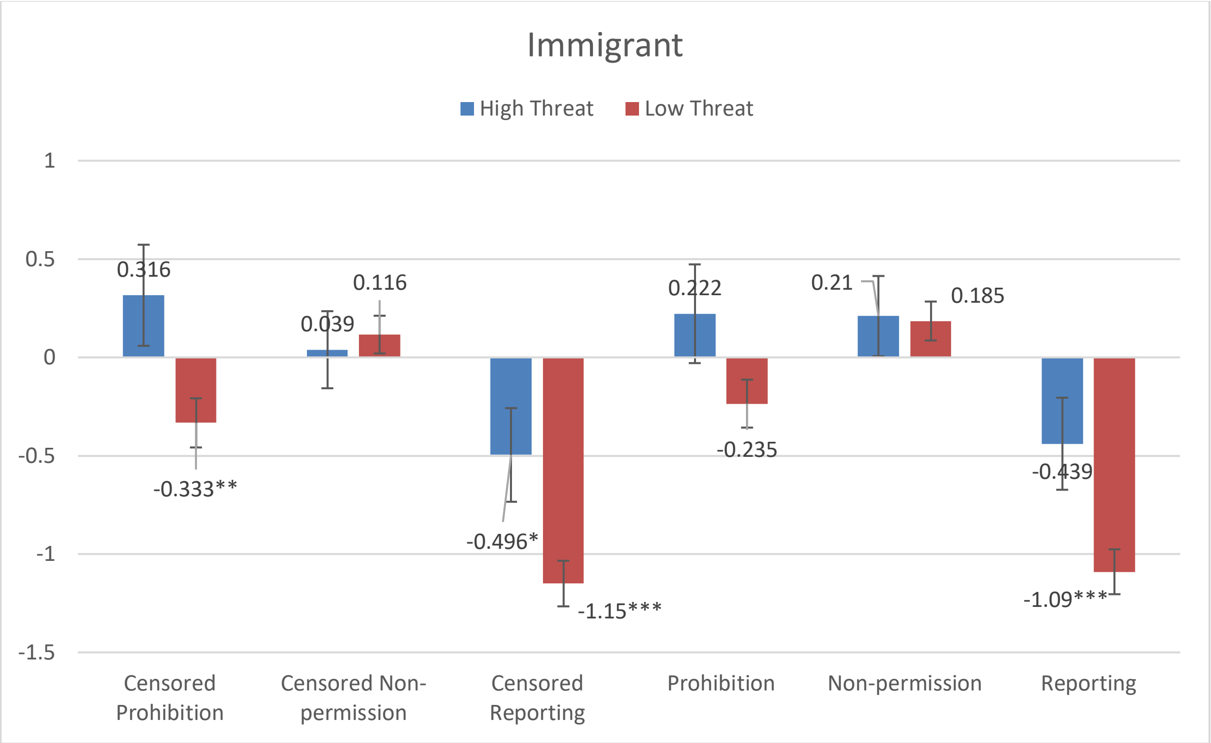
Effects of RWA on prohibition of censored and non-censored hate speech against LGBT, estimated at high and low levels of perceived threat.



Note: This figure demonstrates the parameters estimates (B) from Table 3.4. Parameter estimates for RWA-effects on prohibition of hate speech against LBGT, estimated at high (one SD above the mean) and low (one SD below the mean) levels of perceived threat in the different normativity conditions (censored, non-censored). **p* < 0.05; ***p* < 0.01; ****p* < 0.001. Error bars represent standard errors.

Figure 3.3

Effects of RWA on prohibition of censored and non-censored hate speech against immigrants, estimated at high and low levels of perceived threat.



Note. This figure demonstrates the parameter estimates (B) from Table 3.4. Parameter estimates for RWA-effects on the prohibition of hate speech against immigrants, estimated at high (one SD above the mean) and low (one SD below the mean) levels of perceived threat in the different normativity conditions (censored, non-censored). **p* < 0.05; ***p* < 0.01; ****p* < 0.001. Error bars represent standard errors.

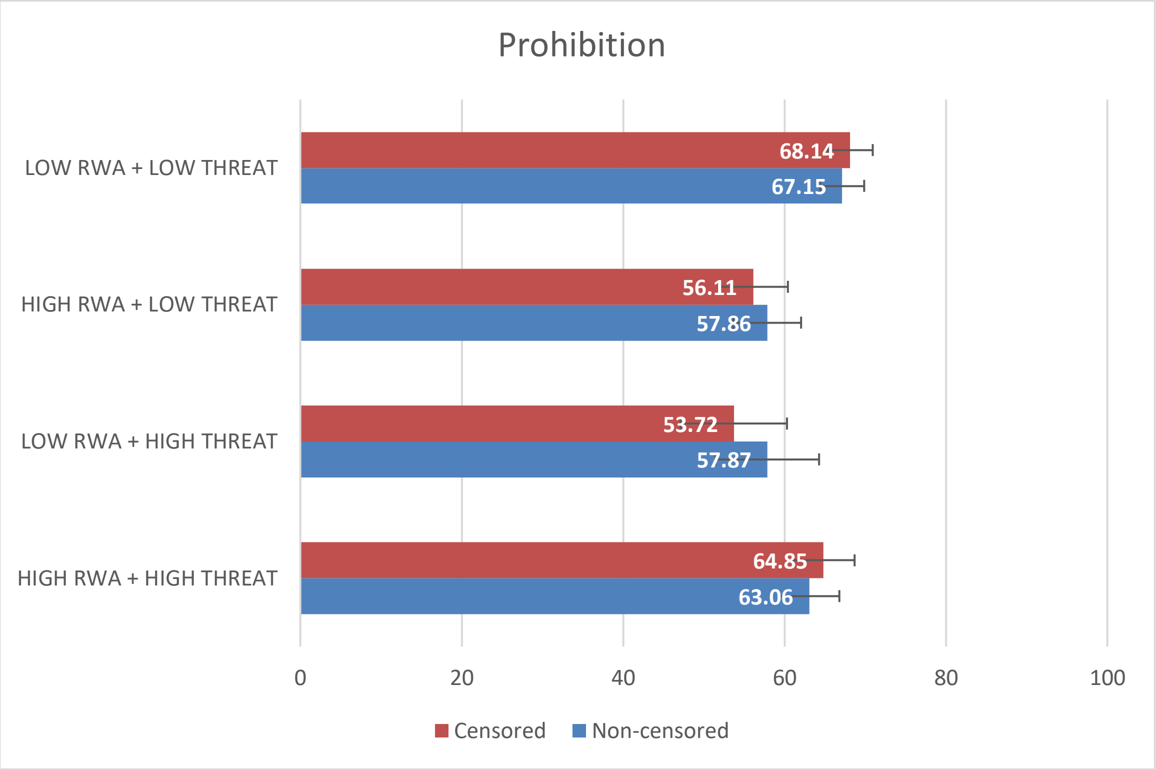
Therefore, we can conclude that the negative relation between RWA and hate-speech prohibition (item 1) and reporting (item 3) is stronger when the target groups are not threatening than when they are threatening, which is clearly in contradiction with hypothesis H3.

In addition, from the analysis of the simple slopes, to interpret the 4-way interaction from a different perspective, we also plotted the estimated marginal means of hate-speech prohibition, adjusted for high and low levels of RWA and perceived threat. In this analysis, we can see the mean of hate speech prohibition (censored and non-censored) for each group of the 4 different combinations of high levels (1 SD above the mean) and low levels (1 SD below the mean) of RWA and perceived threat, estimated at the respective specific values of the covariates (Figures 3.4 – 3.6).

Analysing the prohibition dependent variable with the normativity moderation (Censored x Non-censored) for the 4 different levels of RWA and perceived threat, as can be seen in Figure 3.4, people who support hate speech prohibition more are people with either high levels of RWA combined with high levels of threat or people with low levels of RWA combined with low levels of threat. This interaction seems to be slightly stronger for censored items than for non-censored items. (See Figure 3.4).

Figure 3.4

Support of prohibition of Hate-speech estimated for the 4 different levels of RWA and perceived threat.

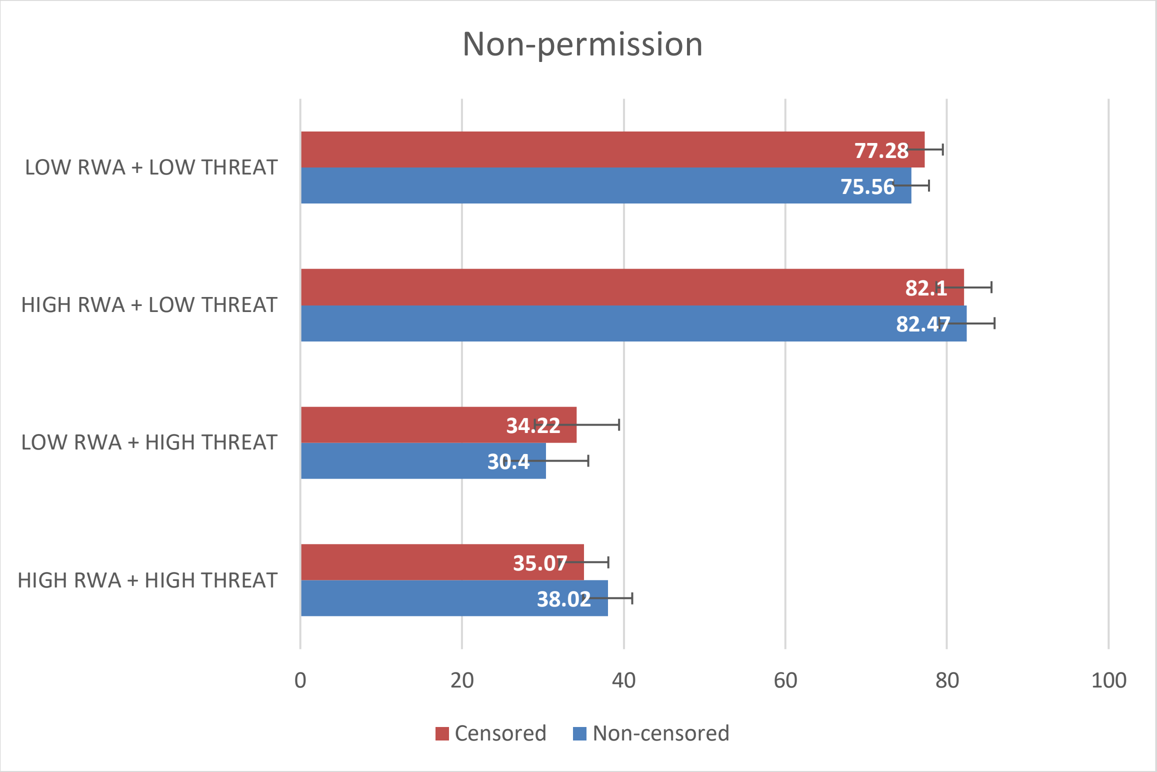


Note: This figure demonstrates the estimated marginal means of the Prohibition dependent variable estimated at high (+1*SD*) and low (-1*SD*) levels of Threat and RWA and in the different normativity conditions. Error bars represent standard errors.

Results were different for the non-permission and reporting items. People who support more the permission of hate speech on the platform (reversed coded as non-permission) were people with high levels of perceived threat, as can be seen in Figure 3.5, whereas RWA did not have much of an impact.

Figure 3.5

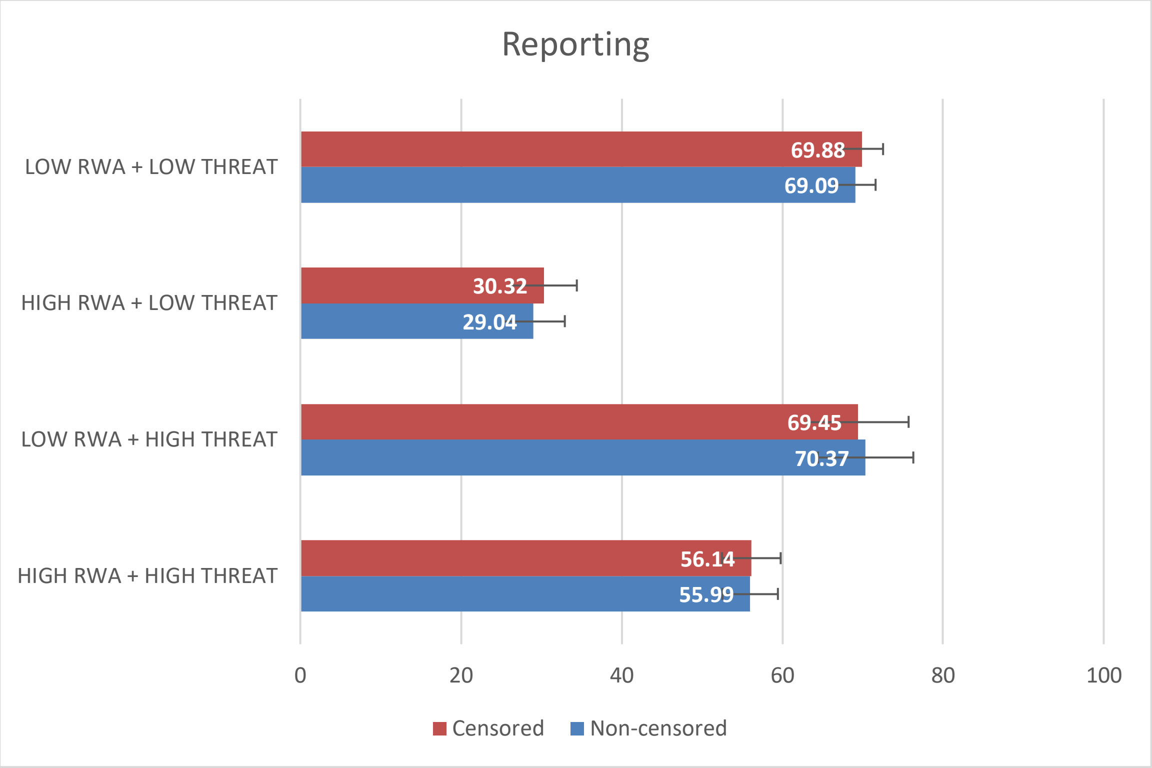
Support non-permission of Hate-speech estimated for the 4 different levels of RWA and perceived threat



Note: This figure demonstrates the estimated marginal means of the Non-permission dependent variable estimated at high (+1*SD*) and low (-1*SD*) levels of Threat and RWA and in the different normativity conditions. Error bars represent standard errors.

Figure 3.6

Support of reporting of Hate-speech estimated for the 4 different levels of RWA and perceived threat.



Note: This figure demonstrates the estimated marginal means of the Reporting dependent variable estimated at high (+1*SD*) and low (-1*SD*) levels of Threat and RWA and in the different normativity conditions. Error bars represent standard errors.

Regarding the support for reporting hate speech, people with high levels of RWA would report less hate speech, when compared to those with low levels of RWA, and this effect is stronger for low than for high levels of threat (Figure 3.6.)

Table 3.4

Parameter Estimates of RWA effects, estimated at high (1SD above the mean) and low (1SD below the mean) levels of perceived threat. Estimates taken from the GLM with hate speech prohibition items as dependent variable, normativity (censored vs. non-censored) as within-subjects factor, and RWA and perceived threat as continuous predictors (all interactions included in the model).

						95% Confidence Interval		
Dependent Variable	Parameter	*B*	Std. Error	*t*	*p.*	Lower Bound	Upper Bound	Partial Eta Squared
At High Levels of Threat (1*SD* above the Mean)								
IC1	RWA	0.316	0.257	1.233	.219	-0.189	0.822	.008
IC2_r	RWA	0.039	0.196	0.200	.842	-0.348	0.426	.000
IC3	RWA	-0.496	0.238	-2.08	.039	-0.966	-0.026	.021
IS1	RWA	0.222	0.251	0.886	.377	-0.273	0.717	.004
IS2_r	RWA	0.210	0.204	1.02	.305	-0.193	0.612	.005
IS3	RWA	-0.439	0.234	-1.87	.063	-0.901	0.024	.017
HC1	RWA	0.304	0.260	1.16	.244	-0.209	0.818	.007
HC2_r	RWA	0.009	0.211	0.040	.968	-0.407	0.424	.000
HC3	RWA	-0.246	0.249	-0.988	.325	-0.738	0.245	.005
HS1	RWA	0.067	0.256	0.260	.795	-0.439	0.572	.000
HS2_r	RWA	0.215	0.204	1.054	.293	-0.187	0.617	.005
HS3	RWA	-0.363	0.232	-1.56	.119	-0.820	0.094	.012
At Low Levels of Threat (1*SD* below the Mean)								
IC1	RWA	-0.333	0.125	-2.66	.008	-0.579	-0.087	.034
IC2_r	RWA	0.116	0.096	1.21	.224	-0.072	0.305	.007
IC3	**RWA**	**-1.15**	**0.116**	**-9.93**	**<.001**	**-1.38**	**-0.923**	**.329**
IS1	RWA	-0.235	0.122	-1.92	.056	-0.476	0.006	.018
IS2_r	RWA	0.185	0.099	1.86	.064	-.011	.381	.017
IS3	**RWA**	**-1.09**	**0.114**	**-9.62**	**<.001**	**-1.324**	**-.874**	**.316**
HC1	RWA	-0.338	0.127	-2.666	.008	-.588	-.088	.034
HC2_r	RWA	0.152	0.103	1.48	.140	-.050	.355	.011
HC3	**RWA**	**-1.052**	**0.121**	**-8.66**	**<.001**	**-1.291**	**-.813**	**.272**
HS1	RWA	-0.283	0.125	-2.268	.024	-.529	-.037	.025
HS2_r	RWA	0.200	0.099	2.01	.045	.004	.395	.020
HS3	**RWA**	**-1.13**	**0.113**	**-10.0**	**<.001**	**-1.356**	**-.911**	**.334**

Note. *P values* below .001 are in bold
IC1= Immigrant Censored Hate Speech Prohibition; IC2_r = Immigrant Censored Hate Speech Non-Permission; IC3 = Immigrant Censored Hate Speech Report; IS1 = Immigrant Non-Censored Hate Speech Prohibition; IS2_r = Immigrant Non-Censored Hate Speech Non-Permission; IS3 = Immigrant Non-Censored Hate Speech Report; HC1 = LGBT Censored

Hate Speech Prohibition; HC2_r = LGBT Censored Hate Speech Non-Permission; HC3 = LGBT Censored Hate Speech Report; HS1 = LGBT Non-Censored Hate Speech Prohibition; HS2_r = LGBT Non-Censored Hate Speech Non-Permission; HS3 = LGBT Non-Censored Hate Speech Report

To sum up the results and additional analysis we summarized the most important effects in Table 3.5.

Table 3. 5

Items	Research Hypotheses	*df(Error)*	*F*	*P*	η_p^2	Results
H1	RWA main effect	**1.000(199)**	**60.2**	**<.001**	**.232**	Rejected[1]
H2	Normativity * RWA	1.000(199)	5.71	.018	.028	Partially Accepted[2]
H3	RWA * Manipulation Condition *Target-group	2.000(199)	0.610	.545	.006	No manipulation effect
Additional Analyses	Perceived Threat * Items	**1.719(345.455)**	**84.38**	**<.001**	**.296**	Unpredicted[3]
Additional Analyses	RWA* Perceived Threat	1.000(201)	6.56	.011	.032	Unpredicted
Additional Analyses	RWA*Items*Perceived Threat	1.719(345.455)	4.07	0.18	.020	Unpredicted[4]
Additional Analyses	RWA*Perceived Threat*Items* Normativity	1.967(401.202)	6.84	.001	.033	Unpredicted

Note. H1 = Hypothesis one; H2 = Hypothesis two; H3= Hypothesis tree. *Note. P values* below .001 are in bold

[1] = Although the main effect of RWA was significant when analyzing the parameter estimates the coefficients (*B*) for RWA effects on all three dependent variables were negative (Table 3.3), different from what we expected.

[2] = The moderation is partially accepted because, although there was no positive relation between RWA and hate speech prohibition, the unpredicted negative relation with normativity was weaker for censored expressions.

[3]= Perceived threat was negatively related to non-permission, positively related to reporting, and unrelated to prohibition (Figure 3.1).

[4]= The parameter estimates analysis revealed that under high threat (1 *SD* above the mean; Table 3.4, upper part) the main effect of right-wing authoritarianism became positive and non-significant for items 1 (prohibition) and 2 (non-permission). For item 3 (reporting) the

parameter of the RWA effect remained negative. For participants with a low perceived threat (1 *SD* below the mean; Table 3.4, lower part), the main effect of RWA was again positive and mostly non-significant (only significant for non-censored condition in LGBT as target) for item 2 (non-permission) but remained negative and mostly significant for item 1 (prohibition) and item 3 (reporting).

CHAPTER 4

Discussion

Hate speech continues to be a central topic on the political agenda in Portugal and in the world. Intending to combat this practice, it is crucial to understand how this phenomenon happens and manifests in the virtual and real environment. Building on previous research regarding hate speech and right-wing authoritarianism (e.g Bilewicz et al., 2015), behavior online (e.g Álvarez-Benjumea & Winter, 2018), and perceived threat (e.g Cohrs & Asbrock 2009), this study investigated whether ideological beliefs would affect the support of hate speech prohibition and if perceived threat and social norms would moderate these effects.

The present study examined the effect of authoritarianism on the support of online hate speech prohibition. More precisely, replicating previous research, the current study tested if authoritarianism predicts confrontation against hate speech on social media. Thus, we conducted a study investigating whether right-wing authoritarianism (I.V) can positively predict support of the prohibition, non-permission, and reporting of hate speech (Hypothesis one). Previous research had found such a relationship (Bilewicz et al.,2015; Bilewicz et al. 2020). Going beyond previous research we added to the study the test of moderating effects of threat and social norms.

We predicted that the normativity of hate speech moderates the relation between right-wing authoritarianism and hate speech prohibition, with a stronger positive relation when hate speech is non-normative as compared to when it is normative (Hypothesis two). This hypothesis was derived from the reasoning that the positive relation between RWA and hate-speech prohibition found in previous research could be explained by the authoritarianism component of strong adherence of social norms and aversion against norm violations. We also predicted that intergroup threat would moderate the relation between RWA and hate speech prohibition as well, with a more positive relation under low threat than under high threat (Hypothesis three). The reasoning behind this hypothesis was that the known positive relation between RWA and prejudice, which is stronger for groups that are perceived as threatening (Cohrs & Asbrock (2009), should counteract the positive relation between RWA and hate-speech prohibition.

The hypotheses were tested in an experimental survey. Data was collected via a Qualtrics questionnaire. As a manipulation of intergroup threat, participants were randomly assigned to one of three experimental conditions (between-subjects). They were presented with two newspaper articles, about either the LGBT community or immigrants in Portugal or about an unrelated topic (control condition). After finishing reading the newspaper articles, participants filled in the RWA Scale, Social Distance Scale, and Intergroup Threat Scale for both target groups.

Following these measures, we presented eight examples of online hate speeches against members of the LGBT community and eight examples of online hate speeches against immigrants. We asked participants whether they would permit, prohibit, or report these online hate speech posts. In this phase of the study, we manipulated normativity, so in some hate speech publications, we added a warning saying that the publication was removed from the platform. We presented these censored publications randomly to participants (within-subject manipulation). Each post was never seen more than once. In total, participants read 16 hate speeches (eight for LGBT, four censored and four non-censored, and eight for immigrants, four censored and four non-censored).

Because hate speech is considered an uncivil behavior by being deviant from the norm, we expected that individuals with high levels on the RWA scale, which are usually people that prefer strict rules and social control, would be more likely to confront online hate speech. Instead, in our research, the effect of authoritarianism on supporting hate speech prohibition was negative. Thus, in our research, authoritarianism did not prevent hate speech online. Our results also show that the impact of authoritarianism depends on the way how opposition to hate speech is measured. The RWA had the strongest negative relation with the reporting of hate speech, and the weakest negative relation with the item directly measuring the prohibition of hate speech, whereas the strength of the negative relation of authoritarianism with not-permitting hate speech (i.e., the positive relation with permitting hate speech) was in-between these extremes.

We also assumed that the degree of normativity would affect the relation between authoritarianism and support for the prohibition of hate speech (Hypothesis H2). We expected that introducing a censoring warning attached to the hate-speech post (“deleted from the platform”) would reinforce authoritarians’ aversion to norm violation and increase their willingness to act against hate speech. Regarding our results, we consider H2 partially supported, since the interaction of authoritarianism with censoring hate speech was significant. Even though the relation between RWA and the prohibition of hate-speech was negative (contrary to previous results), this negative relation was overall weaker for censored than for non-censored hate speech. The results were slightly more homogeneous across different items for hate-speech against LGBT than against immigrants. Nevertheless, this interaction between normativity and RWA was not qualified by interactions with target group and/or item, and overall, it goes in the direction that one would expect if one assumed that the explanation for the positive relation of RWA with hate speech prohibition found in previous research lies in authoritarian's aversion to norm violations.

The experimental between-subjects manipulation that we introduced to test the moderation of RWA effects by perceived intergroup threat (Hypothesis H3) was not successful, and it is therefore no surprise that there was no significant interaction between RWA and the threat manipulation either. For that reason, we decided to test H3 in an additional analysis by using measured perceived threat as a moderator variable rather than the manipulated threat conditions. Here, for this analysis, we included the centered continuous measure of intergroup threat (collapsed across target groups) and its interactions as a factors in the model and afterwards conducted simple slopes analysis estimating the RWA effects on support for hate speech prohibition at high (one standard deviation above the mean) and low (one standard deviation below the mean) levels of perceived threat.

We expected that authoritarians who felt threatened would lose control of their aversion against the target group, which would overweight their aversion against the norm-violation that hate speech implies and which they should usually have because of their preference for strict rules and social control. As a result, they would become less supportive of the prohibition of hate speech. Although we found a significant interaction between RWA and perceived threat, this interaction was not consistent across items and normativity conditions and overall did not support H3 in any way. In cases in which this interaction occurred the effect was in the opposite direction of what H3 would have predicted.

Interestingly, some unexpected results such as the main effect of perceived threat in prohibiting hate speech, which interacted with item, suggest that perceived threat increased the willingness to permit hate speech against both target groups, independent of normativity, but also the willingness to report hate speech. So, it seems that in our study threat was related to the belief that hate speech should remain on the platform and should be reported but should not be prohibited. This suggests that perceived threat might be linked to the belief that hate speech needs to be analyzed according to the platform's policies.

How to explain such ambivalent response to threat? On one hand, the willingness to permit hate speech when one feels threatened could have several reasons, such as agreement with the hate speech content or identification with the hate speech's intention to act against the threatening target group. This is somewhat, but not entirely, in line with the theoretical reasoning by Bilewicz et al. (2015) who identified differences in authoritarians' responses to hate speech directed against minorities whom they consider threatening versus the ones whom they consider harmless. They propose that "it is possible that under some circumstances, people high in RWA may tolerate hate speech or even use it as a tool to protect their ingroup.” (p. 10, 2015) However, given that in our data the threat effect on the permission of hate speech was rather independent of RWA, such threat effects probably generalize to lower levels of authoritarianism. Our data might, therefore, be better explained by RWA-unspecific factors on the intergroup level rather than by authoritarians’ specific individual responses that were proposed by Bilewicz et al. (2015) which were adherence to a stronger normative protection of harmless groups and strong negative arousal, or psychological threat triggered by threatening groups.

On the other hand, we have the unpredicted pattern for the reporting item, with threat increasing the tendency to report, which was particularly strong at high levels of RWA. In other words, overall RWA was negatively related with reporting, but less so when the threat level was high. There are two aspects in this unpredicted pattern that must be explained. The first one is the general tendency to report less for participants high on RWA. It seems logical that people with high RWA were supposed to be stronger in reporting, since the tendency to submit to an authority perceived as legitimate could motivate authoritarians to report and thereby pass the decision to ban hate speech to the platform "authorities" and align with its rules. However, in our study, it may also be that the authoritarians do not perceive the platform's policies as legitimate to society, especially considering that our sample was mostly of right-wing participants – at least compared to other samples that usually participate in this kind of research - and our results may have been affected by this. Hence, high-RWA participants might have abstained from reporting and, if any, their preferred response to hate-speech norm violations might have been to take the decision of not permitting or prohibiting into their own hands.

The second unpredicted pattern - the increase in the tendency to report when the target group is threatening - was particularly strong at participants with high levels of RWA up to the point that they approached levels of reporting shown by low-RWA participants. One possible explanation is that when authoritarians feel highly threatened by the target group and generated by a feeling of lack of control, they become in favor of reporting, in a way that they would self-regulate their prejudices to appear unbiased, consequently they will support reports of hate speech.

Research on aversive racism highlights the process of self-regulation in low-prejudiced people and argues that being aware of racial bias increases the feeling of guilt and compunction and consequently reduces their bias, also, this self-regulation is motivated by a desire to appear nonprejudiced to others (Dovidio & Gaertner, 2004). Because of fear of the negative social consequences of appearing prejudiced people not only tend to rationalize their biases, but also use moral standards that motivate them to respond without prejudice (Bamberg & Verkuyten, 2022). We can speculate that, in our findings, the perception of threat of the target group might have triggered self-regulation of prejudice among authoritarians, which led to their support for reporting online hate speech. But why would people high on RWA want to appear nonprejudiced? It may be that – from their perspective - threatening groups (e.g., LGBTQ+ or immigrants) are seen as powerful and, thus, authoritarians would tend to adhere to the social norms that these powerful groups and their allies impose on society. Thus, as a result to this perception of powerful targets being targeted by hate speech, authoritarians are torn between there prejudice against these groups and the obligation to do something about the non-normative hate speech. Thus, even if they tend to let such hate-speech simply pass, they might feel less confident and might not trust their own decisions and, thus, decide to align with the platform rules and escalate the incident of hate speech to the platform authorities by reporting.

Previous research on hate speech showed that threatening hate speech can also lead to change on the intraindividual level and create a sense of responsibility to take action against speech. In other words, it can be that the more hate speech is considered threatening towards a target group, the more people feel personally responsible to take action against that hate speech (Leonhard et al., 2018). Although this research does not fully correspond with ours, in the sense that what was measured was the threat by hate speech and not the threat by the target group, it is nevertheless interesting research on the intention to counterargue against hate speech and how it relates to perceived threat and responsibility feelings.

Overall results indicate that, unlike what was found in previous research reported in the literature (e.g Bilewicz et al., 2015), authoritarianism does not lead to the support of the prevention of hate speech. In addition, one cannot blame intergroup threat for the absence of that relation, because authoritarian's intentions to report and to prohibit hate-speech did not go up when intergroup threat was relatively low.

A plausible explanation for authoritarianism undermining opposition to hate speech on social media might be the perception of hate speech as something that no longer deviates from the norm. Authoritarians, in this study, seem to judge that hate is common and not a deviant behavior, even with the indication of norm violation (censored hate speech). As stated earlier, in the case of censored hate speech (for LGBT groups) there was a reduction in the intention not to report but the authoritarians continued supporting hate speech more than non-authoritarians. This suggests that hate speech can be interpreted as a common behavior and such interpretation could enhance the spread of hate speech among internet users.

The rise of authoritarian leaders can explain why hate speech is no longer viewed as deviant from the norm by authoritarians. When politicians or religious leaders encourage hate speech, and if this becomes more common in the public debate, then the sense of norm might change, and hate speech would not be considered any more as something that is against the norm (Bilewicz et. al. 2020). So, if the authoritarians in our study are used to environments where hate speech is common ("desensitization"), then their perception of hate speech being deviant from the social norm may have changed and this affected the study's outcome (Bilewicz et al. 2020).

Although our results point to different directions than some research reported in the literature (Bilewicz et al.,2015), they are partially consistent with other research. For instance, Wilhelm and colleagues found that RWA was negatively correlated with flagging (reporting) of hate speech against feminist women and sexual minorities (only significant when the hateful comments came from women; Wilhelm at al., 2018) and unrelated to hate speech against refugees (Wilhelm et al 2019).

Also, in their research, they found a significant direct positive effect of individualization foundation (liberal orientation) on flagging intention, but conservative orientations reduced the flagging intention (Wilhelm et al 2018; Wilhelm et al 2019).

In accordance with these findings, it is important to highlight that our research had a significant number of people with right-wing political orientation, which could have led to a more conservative perception of minority groups. It would be interesting to study whether RWA effects on preventing hate speech vary along the political spectrum of participants and how that interacts with the kind of target of the hate speech.

We can conclude that our experimental study brings new insight into whether ideological beliefs would affect the support of hate speech prohibition and if perceived threat and social norms would moderate these effects. However, we would recommend conducting more research about the relation between RWA and the support of hate speech prohibition and the moderation by perceived threat or other possible moderators and mediators.

4.1. Implications for research and practice

There are three important implications of our results. Because our data indicate (even if only on an interindividual level) that RWA increases the intention to permit, not prohibit and not report hate speech against LGBT and immigrants, hate speech is more likely to be seen and accepted in authoritarian environments, such as in right-wing and conservative online communities: social media groups, blogs, and comment sections in right-wing newspapers. Therefore, to expand the fight against hate, legal interventions in these environments are necessary and can probably be effective.

According to our data, the type of normativity significantly affects the interaction between authoritarians and hate speech support. The literature suggests that removing hateful comments (Alvarez-Benjumea & Winter, 2018) or preventing hate speech from becoming "normal" in social networks (Bilewicz et al., 2020) helps combat hate speech online. Our results point in the same direction, indicating that the removal of hateful content (content moderation) is critical to combat online hate.

However, this type of measure does not fully solve the problem effectively, because today there are social platforms where there is no content moderation, and people who want to spread hate end up migrating to these networks. For example, the social platform GAB is an alternative to Twitter, with no content moderation, where posts made by hateful supporters tend to go further within the platform, spread faster, and have a greater reach among network users (Mathew et al., 2019). Moreover, these hateful users are connected and feed much of the content that is generated within the platform (Mathew et al., 2019) helping not to fully solve the problem.

We can speculate that, although the three dependent variables (prohibit, non-permission and report) have been used to measure intentions against online hate speech, it seems evident that each item represents a distinct meaning, so there are semantic differences that so far have been overlooked. The dependent variable was negatively correlated with perceived threat when it was measured as non-permission, that is, with a reversed item asking for permission of hate speech. While it seems reasonable that threat might increase the tendency to permit hate speech, there was no relation of the dependent variable with threat when it was measured directly as prohibiting, and the dependent variable was even positively predicted by threat when it was measured as reporting. We can only speculate that the support of prohibition is a more direct action and might represent a solid opinion against a hateful content, while supporting the permission is a less direct and less solid action against the threatening group that might be more easily justified with the principle of freedom of speech. Surely, report is an action linked to a specific mechanism of social networks and leaves the decision to exclude a content under the platform's policies. This may explain our results, however, it would be interesting to investigate more about those outcomes by exploring (e.g., in qualitative research) the semantic difference among the dependent variables for online users as well as observe their interaction with hate speech comments.

Moreover, further research should take a closer look at authoritarianism as a predictor of hate speech and the moderation by normativity since studies on social norms (Álvarez-Benjumea & Winter, 2018b; Bilewicz & Soral, 2020b; Kunst et al., 2021; Wilhelm & Joeckel, 2019) show that different norms (descriptive and injunctive) influence and have different effects on attitudes and behavior.

Also, for a better understanding, future studies should investigate the influence of justification (such as denying the negative intent of violent comments) as a predictor of supporting hate speech and liberal foundations as a predictor to report hate speech (Wilhelm et al 2019).

4.2. Limitations and future research

Naturally, the present study's limitations should be taken into consideration when discussing our findings. First our manipulation of threat was not successful, thus, the selection of the newspaper did not reverberate how we expected. We tried to simulate participants' experiences and, by doing that, to come as close as possible to the actual use of social networks. Therefore, the manipulation was done using real newspaper articles in a post on Twitter. Contrary to our belief that these newspaper articles could induce threat, this manipulation was not strong enough. For future research, we would suggest using more threatening manipulation, such as artificial material or artificial newspapers with a strong threatening message against a group.

Future studies may investigate if our findings, the unpredicted result that right-wing authoritarianism negatively predicted support of hate speech prohibition and that perceived threat increased the willingness to permit and the willingness to report hate speech, independent of normativity, is replicated in different target groups, for instance religious groups or women.

Another limitation of the study is that we did not categorize the hate speech collected, so we do not know whether the selection of hate speech may have influenced the results. Previous studies have shown that direct offense (i.e., individual person) or a call for violence (i.e., "all mosques should be burned") is more likely to be reported compared to less direct offense (abstracted group, i.e., Muslims, immigrants) or more disguised or subtle norm violation (i.e., political agitation, spreading rumors or conspiracies, and defamation – "immigrants are social parasites") (Wilhelm et al 2019). Also, there is research showing that threatening hate speech creates a sense of responsibility to take action against such speech. (Leonhard et al., 2018). For future research, we would distinguish the types of norm violations in hate comments for better control of the outcomes.

We may also only speculate whether our data would replicate in another cultural context. Our study is not culture limited since the issue we addressed (hate speech against LGBT and immigrants) happens globally. It can be argued that the Portuguese authoritarians may have been influenced by the Portuguese cultural context since Portugal's censorship was common for many years in the previous century and there is a strong aversion among Portuguese people to prohibit or limit speech (Gonçalves et. al., 2021). But it is unlikely that this cultural context interfered with our research, as it shows that Portuguese people who score low on the RWA scale are in favor of the non-permission of hate speech, prohibit it and report it. Nevertheless, it is recommended to replicate the study in different social and cultural contexts.

In sum, this exploratory study investigated the influence of political ideology on the prohibition of hate speech. Our results show that authoritarians do not prevent hate speech more than non-authoritarians do and that perceived threat increases the willingness to permit and the willingness to report hate speech in authoritarians, independently from normativity of hate speech.

In general, despite the limitations presented above, our study brings new ideas to the investigated field, and these new insights are especially relevant for research in the area, being an important material for the elaboration of strategies to combat the spread of hate speech.

References

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality*. Harpers & Brothers.

Allport, G. W. (1954). *The nature of prejudice*. Double-day Anchor.

Altemeyer, B. (1998). The otis her "authoritarian personality". Advances in Experimental Social Psychology. *Academic Press*, *v*(30), 47-91. https://doi.org/10.1016/S0065-2601(08)60382-2

Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237. https://doi:10.1093/esr/jcy005

Article 19 (Organization). (2012). *Prohibiting incitement to discrimination, hostility or violence*. Article 19. https://www.refworld.org/pdfid/50bf56ee2.pdf

Bamberg K., & Verkuyten M., (2022) Internal and external motivation to respond without prejudice: a person-centered approach. *The Journal of Social Psychology*, *162*(4), 435-454. https//10.1080/00224545.2021.1917498

Barendt, E. (2019). What is the harm of hate speech? *Ethical Theory and Moral Practice*, *22*(3), 539–553. https://doi.org/10.1007/s10677-019-10002-0

Bayer, J., & Bárd, P. (2020). *Hate speech and hate crime in the EU and the evaluation of online content regulation approaches.* European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Bilewicz, M., S, W., Marchlewska, M., & Winiewski, M. (2015). When authoritarians confront prejudice: Differential effects of SDO and RWA on support for hate-speech prohibition. *Political Psychology*, *38*(1), 87–99. https://doi.org/10.1111/pops.12313

Bizumic, B., & Duckitt, J. (2018). Investigating right wing authoritarianism with a very short authoritarianism scale. *Journal of Social and Political Psychology*, *6*(1), 129–150. https://doi:10.5964/JSPP.V6I1.835

Boeckmann, R. J., & Turpin-Petrosino, C. (2002). Understanding the harm of hate crime. *Journal of Social Issues*, *58*(2), 207–225. https://doi.org/10.1111/1540-4560.00257

Casa do Brasil de Lisboa. (2021). *Discurso de ódio e imigração em Portugal: Diagnóstico do projeto #migramyths – Dismistificando a imigração.* https://casadobrasildelisboa.pt/discurso-de-odio-e-imigracao-em-portugal-diagnostico-do-projeto-migramyths-desmistificando-a-imigracao-2a-edicao/

Carlson, C., R., (2021). *Hate Speech (The MIT Press Essential Knowledge series).* MIT Press.

Carnaghi, A., & Maass, A. (2008). Derogatory language in intergroup context: Are "gay" and "fag" synonymous? In Y. Kashima, K. Fiedler, & P. Freytag (Eds.), *Stereotype dynamics: Language-based approaches to the formation, maintenance, and transformation of stereotypes*, (pp. 117–134). Lawrence Erlbaum Associates Publishers.

Chang, R., & Little, T. D. (2018). Innovations for evaluation research: Multiform protocols, visual analog scaling, and the retrospective pretest–posttest design. *Evaluation & The Health Professions*, *41*(2), 246–269. https://doi:10.1177/0163278718759396

Cohrs, J. C., & Asbrock, F. (2009). Right-wing authoritarianism, social dominance orientation and prejudice against threatening and competitive ethnic groups. *European Journal of Social Psychology*, *39*(2), 270–289. https://doi.org/10.1002/ejsp.545

Crandall, C. S., & Eshleman, A. (2003). A justification – suppression model of the expression and experience of prejudice. *Psychological Bulletin*, *129*(3), 414 – 446. https://doi.org/10.1037/0033-2909.129.3.414

Crandall, C. S., Eshleman, A., & Brien, L. O. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359 – 378. https://doi:10.1037//0022-3514.82.3.359

Decreto de Lei no 65/98 de 2 de Setembro da Assembleia da República. Diário da República, N. 202 (1998). https://dre.pt/dre/home

Doty, R. M., Peterson, B. E., Winter, D. G., Brown, R., Burnstein, E., Manis, M., & Stewart, A. J. (1991). Threat and authoritarianism in the united states, *Journal of Personality and Social Psychology, 61*(4), 629–640. https://doi:10.1037//0022-3514.61.4.629

Dovidio, J.F., & Gaertner, S.L. (2004). Aversive racism. In M.P. Zanna (Ed.), *Advances in experiemntal social psychology* (vol. 36, pp. 1052). Academic Press

Duckitt, J. (2020). Authoritarianism: Conceptualization, research, and new developments. In Forthcoming in: G.G. Sibley & D. Osborne (Eds). *The Cambridge Handbook of Political Psychology*. Cambridge University Press.

Duckitt, J., Bizumic, B., Krauss, S. W., & Heled, E. (2010). A tripartite approach to right-wing authoritarianism: The authoritarianism-conservatism-traditionalism model. *Political Psychology, 31*(5), 685–715. https://doi.org/10.1111/j.1467-9221.2010.00781.x

Duckitt, J., & Sibley, C. (2007). Right-wing authoritarianism, social dominance orientation, and the dimensions of generalized prejudice. *European Journal of Personality, 21*(2), 113–130. https://doi:10.1002/per.614

Duckitt, J., & Sibley, C. (2010). Personality, ideology, prejudice, and politics: A dual-process motivational model. *Journal of Personality, 78*(6), 1861-93. https://doi:10.1111/j.1467-6494.2010.00672.x

Duckitt, J., & Sibley, C. G. (2016). Personality, ideological attitudes, and group identity as predictors of political behavior in majority and minority ethnic groups. *Political Psychology, 37*(1), 109–124. https://doi:10.1111/pops.12222

ECRI. (2016). *On combating hate speech*. Council of Europe. https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-com-bating-hate-speech/16808b5b01

ESS Round 7: European Social Survey Round 7 Data (2014). *Data file edition 2.2. Sikt - Norwegian Agency for Shared Services in Education and Research, Norway – Data Archive and distributor of ESS data for ESS ERIC*. https:doi:10.21338/NSD-ESS7-2014

European Commission Brussels. (2018). Flash eurobarometer 469 (illegal content online). *GESIS Data Archive*, ZA6962 Data file Version 1.0.0, https://doi.org/10.4232/1.13147

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. United Nations Educational, Scientific and Cultural Organization.

Gonçalves, J., Weber, I., Masullo, G., Torres da Silva, M. & Hofhuis, J. (2021). Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. *New Media and Society*, 1–23. SAGE. https://doi:10.1177/14614448211032310

Gottfredson, M. R., & Hirschi, T. (1987). The methodological adequacy of longitudinal research on crime. Criminology, 25, 581-614. doi:10.1111/j.1745-9125.1987.tb00812.x

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2010). *Multivariate data analysis*. Prentice Hall.

ILGA Europe. (n.d.). *Hate crime & hate speech*. https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech.

Kaplin, W. A. (2016). A proposed process for managing the first amendment aspects of campus hate speech. *The Journal of Higher Education, 63*(5), 517-538. https://doi.org/10.1080/00221546.1992.11778387

Kunst, M., Porten-Cheé, P., Emmer, M., & Eilders, C. (2021b). Do "Good Citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology &Amp; Politics*, *18*(3), 258–273. https://doi.org/10.1080/19331681.2020.1871149

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

Mather, D. M., Jones, S. W., & Moats, S. (2017). Improving upon bogardus: Creating a more sensitive and dynamic social distance scale. *Survey Practice, 10*(4), 1–9. https://doi:10.29115/sp-2017-0026

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019, June 26). Spread of Hate Speech in Online Social Media. *Proceedings of the 10th ACM Conference on Web Science*. https://doi.org/10.1145/3292522.3326034

OHCHR. (2018). *Report of the independent international fact-finding mission on Myanmar*. Human Rights Council, 39 session, 10-28, Agenda item 4. https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action

Onraet, E., Dhont, K., & Van Hiel, A. (2014). The Relationships Between Internal and External Threats and Right-Wing Attitudes. *Personality and Social Psychology Bulletin*, *40*(6), 712–725. https://doi.org/10.1177/0146167214524256

Sibley, C., & Duckitt, J. (2008). Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review, 12*(3), 248-279. https://doi:10.1177/1088868308319226

Stephan, W. G., Ybarra, O., & Bachman, G. (1999), Prejudice toward immigrants. *Journal of Applied Social Psychology, 29*(11), 2221–2237. https://doi:10.1111/j.1559-1816.1999.tb00107.x

Tynes, B. M., Rose, C. A., & Markoe, S. L. (2013). Extending campus life to the internet: Social media, discrimination, and perceptions of racial climate. *Journal of Diversity in Higher Education, 6*(2), 102–114. https://doi:10.1037/a0033267

Tjipto, S., Haksi Mayawati, E., & Bernardo, A. B. (2019). Perceived threat of homosexual in indonesia: Construct, measurement, and correlates. *Makara Human Behavior Studies in Asia, 23*(2), 181-193. https://doi:10.7454/hubs.asia.1111219

UNESCO. (2022, June 16). *What you need to know about hate speech.* https://www.unesco.org/en/countering-hate-speech/need-know

Vallejo-Martín, M., Canto, J.M., San Martín García, J.E., & Perles Novas, F. Prejudice towards immigrants: The importance of social context, ideological postulates, and perception of outgroup threat. *Sustainability*, *13*(9), 4993. https://doi.org/10.3390/su13094993

Willis-Esqueda, C., Delgado, R. H., & Pedroza, K. (2017). Patriotism and the impact on perceived threat and immigration attitudes. *Journal of Social Psychology, 157*(1), 114–125. https://doi:10.1080/00224545.2016.1184125

Windisch, S., Wiedlitzka, S., & Olaghere, A. (2021). PROTOCOL: Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews, 17*(1). https://doi:/10.1002/cl2.1133

Wilhelm, C., Joeckel, S., & Ziegler, I. (2019b, June 22). Reporting Hate Comments: Investigating the Effects of Deviance Characteristics, Neutralization Strategies, and Users' Moral Orientation. *Communication Research*, *47*(6), 921–944. https://doi.org/10.1177/0093650219855330

Wilhelm, C., & Joeckel, S. (2018, July 31). Gendered Morality and Backlash Effects in Online Discussions: An Experimental Study on How Users Respond to Hate Speech Comments Against Women and Sexual Minorities. *Sex Roles*, *80*(7–8), 381–392. https://doi.org/10.1007/s11199-018-0941-5

Appendix A

Appendix A - Reliability of the measures

Table 1

Cronbach's alpha from measures

Scale	*M*	*S.D*	*Cronbach's alpha*
Threat_LGBT	156.88	112.84	.96
Threat_Immigrant	147.34	113.95	.95
Social Distance_LGBT	105.20	86.35	.90
Social Distance_Immigrant	106.68	88.31	.95
RWA	284.57	107.66	.614
Hate Speech	-	-	-
Prohibition (overall)			
Immigrant Censored	181.88	60.09	.39
LGBT Censored	184.32	62.26	.47
Immigrant Non-Censored	182.70	59.52	.41
LGBT Non-Censored	183.82	60.60	.44

Table 2

Factor Matrix^{a} Factorial Analysis of LGBT Threat Scale.

	Factor
	1
Os homossexuais... - ...prejudicam os costumes, as tradições e a vida cultural em Portugal	**.952**
Os homossexuais... - ... possuem valores e crenças que representam uma ameaça às questões morais e religiosas em nossa sociedade	**.921**

Os homossexuais... - ...contribuem para a diminuição da população	**.853**
Os homossexuais... - ...contribuem para a diminuição da população	**.741**

*Note. N = 205. The e*xtraction method was maximum likelihood with an oblimin (Promax with Kaiser normalization) rotation. Factor loading above .30 are in bold.

Table 3

Total Variance Explained of LGBT Threat Scale.

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	**3.259**	**81.479**	**81.479**	3.030	75.752	75.752
2	**.390**	**9.756**	**91.235**			
3	.233	5.813	97.048			
4	.118	2.952	100.000			

Note. Extraction Method: Maximum Likelihood.

Table 4

Item-Total Statistics Very Short Authoritarianism mean scores (M) and standard deviation (SD) for individual items, corrected item-total correlation, and internal consistency (Cronbach's).

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
1. It's great that many young people today are prepared to defy authority. (R*) (É ótimo que muitos jovens hoje estejam preparados para desafiar a autoridade.)	236.283	10720.50	.010	.379	.685

2. What our country needs most is discipline, with everyone following our leaders in unity. (O que nosso país mais necessita é de disciplina, com todos a seguir os nossos líderes em unida.)	231.390	7403.98	.587	.742	.463
3. God’s laws about abortion, pornography, and marriage must be strictly followed before it is too late. (A lei de Deus sobre aborto, pornografia e casamento tem de ser estritamente seguida antes que seja tarde demais.)	239.741	8341.00	.365	.664	.563
4. There is nothing wrong with premarital sexual intercourse. (R*) (NÃO há nada de errado com a relação sexual antes do casamento.)	249.844	8528.11	.371	.333	.560
5. Our society does NOT need tougher government and stricter laws. (R*) (A nossa sociedade NÃO precisa de um governo mais rigoroso e de leis mais rígidas.)	237.356	8907.81	.319	.358	.581

6. The facts on crime and the recent public disorders show we have to crack down harder on troublemakers, if we are going preserve law and order. (Os fatos sobre o crime e as recentes desordens públicas mostram que temos que reprimir mais duramente os criadores de problemas, se quisermos preservar a lei e a ordem.)	228.239	8122.06	.459	.599	.524

Table 5

Descriptive Statistics

	Mean	Std. Deviation	N
RWA	47.4285	17.94414	205
Th_tot	76.0573	54.73025	205
SD_total	70.6309	56.27575	205
IC1	63.7864	27.40271	205
HC1	63.3008	27.94294	205
IS1	62.9648	26.48972	205
HS1	63.5538	27.07885	205
IC3	59.5091	30.22823	205
HC3	59.8137	30.43074	205
IS3	59.3593	29.43373	205
HS3	59.1736	29.50951	205
IC2_r	56.2983	29.95394	205
HC2_r	57.0674	30.53942	205
IS2_r	57.1152	29.75469	205
HS2_r	56.2845	29.68533	205

Note.

IC1= Immigrant Censored Hate Speech Prohibition; IC2_r = Immigrant Censored Hate Speech Non-Permission; IC3 = Immigrant Censored Hate Speech Report; IS1 = Immigrant Non-Censored Hate Speech Prohibition; IS2_r = Immigrant Non-Censored Hate Speech Non-

Permission; IS3 = Immigrant Non-Censored Hate Speech Report; HC1 = LGBT Censored Hate Speech Prohibition; HC2_r = LGBT Censored Hate Speech Non-Permission; HC3 = LGBT Censored Hate Speech Report; HS1 = LGBT Non-Censored Hate Speech Prohibition; HS2_r = LGBT Non-Censored Hate Speech Non-Permission; HS3 = LGBT Non-Censored Hate Speech Report

Table 6

Descriptive statistics and Pearson's correlations.

		RWA	Th_tot	SD_total	IC1	HC1	IS1	HS1	IC3	HC3
RWA	Pearson Correlation	--								
	N	205								
Th_tot	Pearson Correlation	.492**	--							
	Sig. (2-tailed)	<.001								
	N	205	205							
SD_total	Pearson Correlation	.429**	.878**	--						
	Sig. (2-tailed)	<.001	<.001							
	N	205	205	205						
IC1	Pearson Correlation	-.169*	-.093	-.084	--					
	Sig. (2-tailed)	.015	.185	.231						
	N	205	205	205	205					
HC1	Pearson Correlation	-.191**	-.138*	-.108	.849**	--				
	Sig. (2-tailed)	.006	.049	.122	<.001					
	N	205	205	205	205	205				
IS1	Pearson Correlation	-.136	-.094	-.112	.891**	.854**	--			
	Sig. (2-tailed)	.051	.179	.109	<.001	<.001				

	N	205	205	205	205	205	205			
HS1	Pearson Correlation	-.167*	-.097	-.054	.810**	.928**	.817**	--		
	Sig. (2-tailed)	.017	.165	.446	<.001	<.001	<.001			
	N	205	205	205	205	205	205	205		
IC3	Pearson Correlation	-.509**	-.059	-.052	.421**	.423**	.408**	.391**	--	
	Sig. (2-tailed)	<.001	.402	.458	<.001	<.001	<.001	<.001		
	N	205	205	205	205	205	205	205	205	
HC3	Pearson Correlation	-.460**	-.076	-.085	.414**	.467**	.416**	.432**	.906**	--
	Sig. (2-tailed)	<.001	.278	.227	<.001	<.001	<.001	<.001	<.001	
	N	205	205	205	205	205	205	205	205	205
IS3	Pearson Correlation	-.496**	-.058	-.070	.433**	.419**	.432**	.408**	.937**	.924**
	Sig. (2-tailed)	<.001	.407	.322	<.001	<.001	<.001	<.001	<.001	<.001
	N	205	205	205	205	205	205	205	205	205
HS3	Pearson Correlation	-.497**	-.043	-.055	.428**	.450**	.436**	.425**	.875**	.922**
	Sig. (2-tailed)	<.001	.541	.432	<.001	<.001	<.001	<.001	<.001	<.001
	N	205	205	205	205	205	205	205	205	205
IC2_r	Pearson Correlation	-.312**	-.731**	-.687**	.236**	.273**	.283**	.243**	-.143*	-.097
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.041	.165
	N	205	205	205	205	205	205	205	205	205
HC2_r	Pearson Correlation	-.282**	-.694**	-.661**	.255**	.340**	.312**	.298**	-.118	-.087
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.092	.213
	N	205	205	205	205	205	205	205	205	205
IS2_r	Pearson Correlation	-.257**	-.696**	-.668**	.265**	.346**	.349**	.300**	-.149*	-.076

	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.033	.280
	N	205	205	205	205	205	205	205	205	205
HS2_r	Pearson Correlation	-.250**	-.694**	-.667**	.261**	.304**	.332**	.288**	-.128	-.092
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.067	.188
	N	205	205	205	205	205	205	205	205	205

Descriptive statistics and Pearson's correlations.

		IS3	HS3	IC2_r	HC2_r	IS2_r	HS2_r
RWA	Pearson Correlation						
	N						
Th_tot	Pearson Correlation						
	Sig. (2-tailed)						
	N						
SD_total	Pearson Correlation						
	Sig. (2-tailed)						
	N						
IC1	Pearson Correlation						
	Sig. (2-tailed)						
	N						
HC1	Pearson Correlation						
	Sig. (2-tailed)						
	N						
IS1	Pearson Correlation						

	Sig. (2-tailed)				
	N				
HS1	Pearson Correlation				
	Sig. (2-tailed)				
	N				
IC3	Pearson Correlation				
	Sig. (2-tailed)				
	N				
HC3	Pearson Correlation				
	Sig. (2-tailed)				
	N				
IS3	Pearson Correlation	--			
	Sig. (2-tailed)				
	N	205			
HS3	Pearson Correlation	.887**	--		
	Sig. (2-tailed)	<.001			
	N	205	205		
IC2_r	Pearson Correlation	-.130	-.096	--	
	Sig. (2-tailed)	.063	.170		
	N	205	205	205	
HC2_r	Pearson Correlation	-.103	-.111	.928**	--
	Sig. (2-tailed)	.141	.113	<.001	
	N	205	205	205	205

IS2_r	Pearson Correlation	-.134	-.093	.940**	.930**	--	
	Sig. (2-tailed)	.056	.183	<.001	<.001		
	N	205	205	205	205	205	
HS2_r	Pearson Correlation	-.113	-.108	.912**	.947**	.929**	--
	Sig. (2-tailed)	.107	.122	<.001	<.001	<.001	
	N	205	205	205	205	205	205

Note:

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 7

Hate Speech Statements

	Immigrants
1	Todo mundo sabe que as brasileiras são putas, vem cá só para casar com nossos homens, não sei porque sentem-se ofendidas.
2	Os chineses são como herpes, surgem quando menos espera e se multiplicam aos montes, ainda por cima não pagam impostos nessas lojas de merda, por isso que eu não compro nada nesses sítios.
3	Os chamuças do meu bairro fedem, já não os suporto esses indianos.
4	Os imigrantes em Portugal só trazem doenças.
5	Os imigrantes que vivem às custas de benefícios sociais precisam ser expulsos.
6	Não sou racista, mas se não fosse nós os brasileiros ainda estavam a apanhar banana.
7	Eu sou favorável aos imigrantes cá, já que precisamos de pessoas menos intelectuais para fazer o trabalho que nós não queremos fazer.
8	Morte aos imigrantes! Por uma Europa mais branca.

	LGBT
1	Odeio paneleiros crlh.
2	Eu não tenho nada contra os homossexuais, só não quero é ver isso na rua. Se querem respeito, respeitem. Façam-no na vossa casa.
3	Os gays têm que morrer para acabar com a SIDA.
4	Tenho nojo de bichas, são uma degeneração da humanidade e devem ser curadas.
5	Tem muito verme homossexual no nosso país e eles têm que ir embora.
6	Para de ser tão bicha!
7	Se eu tiver um filho gay, resolvo esse problema lhe dando umas pancadas. Filho meu não vira gay!
8	Eu conheço uma lésbica do trabalho que se ofendeu com um comentário meu e foi reclamar com nosso chefe. Eu não tenho nada a ver com o que ela faz ou deixa de fazer, acho que é nojento e ainda querem respeito, é ridículo.

Appendix B

Appendix B - Questionnaire

Q.1 Nacionalidade
1 – É cidadão português?
Sim /Não

2 – Possui dupla nacionalidade?
Sim/Não/

3 – Se não é cidadão português ou se possui dupla nacionalidade, por favor assinale o país que corresponde a sua nacionalidade:
*Lista de países - Quatrics

4 – Está a viver em Portugal nos últimos 2 anos?
Sim /Não

Q.2 Newspaper article (Manipulation)
Por favor, para a realização do questionário, é importante que leia com atenção as seguintes notícias, prestando atenção nas fotos e enunciados apresentados. Após a leitura será perguntado detalhes sobre as notícias lidas.

Q.2.1 Reading Manipulation check
1 – Já havia lido sobre essa notícia?
Sim /Não
2 – Já havia visto essa notícia nas redes sociais?
Sim/Não/

Q.3 Por favor, peço que responda se concordas totalmente ou discordas totalmente das seguintes afirmações:

0 = *Discordo Totalmente*
100 = *Concordo Totalmente*

	Discordo totalmente	Concordo Totalmente
	0	100

É ótimo que muitos jovens hoje estejam preparados para desafiar a autoridade.	
O que nosso país mais necessita é de disciplina, com todos a seguir os nossos líderes em unidade	
A lei de Deus sobre aborto, pornografia e casamento tem de ser estritamente seguida antes que seja tarde demais.	
NÃO há nada de errado com a relação sexual antes do casamento.	
A nossa sociedade NÃO precisa de um governo mais rigoroso e de leis mais rígidas.	
Os fatos sobre o crime e as recentes desordens públicas mostram que temos que reprimir mais duramente os criadores de problemas, se quisermos preservar a lei e a ordem.	

Q.4 Por favor, peço que responda em que grau se sente incomodado ou não das seguintes afirmações.

0 = *Não me incomodava nada*
100 = *Incomodava-me muito*

Se um imigrante de raça ou grupo étnico diferente da maioria portuguesa...

	Não me incomodava nada	Incomodava-me muito
	0	100
...fosse seu vizinho de rua.		
...fosse nomeado seu chefe.		
...casasse com um familiar próximo		

Q.5 Por favor, peço que responda se concordas totalmente ou discordas totalmente das seguintes afirmações.

0 = *Discordo Totalmente*
100 = *Concordo Totalmente*

Os imigrantes que vêm viver para cá...

Discordo totalmente Concordo totalmente

0 100

...prejudicam os costumes, as tradições e a vida cultural em Portugal.	
...prejudicam as crenças e práticas religiosas em nossa sociedade.	
...são maus para a economia portuguesa.	
...faz com que a criminalidade aumente em Portugal.	

Q.6 Por favor, peço que responda em que grau se sente incomodado ou não das seguintes afirmações.

0 = *Não me incomodava nada*
100 = *Incomodava-me muito*

Se um homossexual...

Não me incomodava nada Incomodava-me muito

0 100

...fosse seu vizinho de rua.	
...fosse nomeado seu chefe.	
...casasse com um familiar próximo	

Q.7 Por favor, peço que responda se concordas totalmente ou discordas totalmente das seguintes afirmações.

0 = *Discordo Totalmente*
100 = *Concordo Totalmente*

Os homossexuais...

	Discordo totalmente	Concordo totalmente
	0	100
...prejudicam os costumes, as tradições e a vida cultural em Portugal		
... possuem valores e crenças que representam uma ameaça às questões morais e religiosas em nossa sociedade		
...contribuem para a diminuição da população		
...aumentam os níveis de doenças sexualmente transmissíveis em nossa sociedade portuguesa		

Q.8 – Por favor, leia cada exemplo de post na rede social e avalie, em uma escala de 0 à 100 se concordas totalmente (100) ou discorda totalmente (0) se cada uma das afirmações apresentadas deveria ser permitida ou proibida nas plataformas de redes sociais.
Alguma dessas afirmações foram removidas pela plataforma e outras não, mas estamos interessados em perceber sua opinião em cada exemplo.

De acordo com sua opinião sobre o post lido:

	Discordo totalmente	Concordo totalmente
	0	100
Acho que deveria ser PROIBIDO essa afirmação das plataformas de redes sociais.		
Acho que deveria ser PERMITIDO essa afirmação das plataformas mesmo que as pessoas não gostem.		
Acho que essa afirmação deveria ser REPORTADAS das plataformas de redes sociais.		

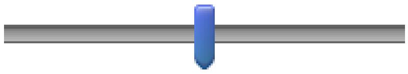
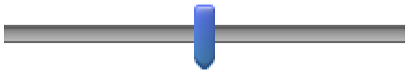
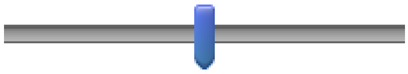
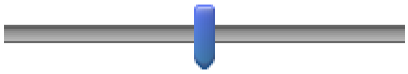
Q.9. Por favor, indique se concorda ou discorda com as seguintes afirmações

Com que frequência utilizas redes sociais (e.x: Facebook, Twitter, ClubHouse, Instagram e etc...)
Uma (1) vez por semana/ Duas (2) vezes por semana/ Três (3) vezes por semana/ Todos os dias.

Q.10 Por favor, leia cada afirmação e em que grau discordas ou concordas com cada afirmação.

Discordo totalmente Concordo Totalmente

0 100

A Internet é segura para os seus utilizadores ()	
É necessário tomar medidas para limitar a disseminação de conteúdo ilegal na Internet ()	
A liberdade de expressão precisa de ser protegida online ()	
Os serviços de alojamento na Internet são eficazes a lidar com conteúdo ilegal ()	

Q. 11 Demographics

Por favor, indique o seu género:

- Feminino
- Masculino
- Outro
- Não gostaria de responder

11.2 Quantos anos tens:

- Qualtrics List.

11.3 Como você se identifica?

- Heterossexual
- Homossexual
- Bissexual
- Outro
- Prefere não dizer

11.4 Qual o seu nível de educação?

- 3 Ciclo
- Secundário
- Licenciatura
- Mestrado
- Doutoramento

Q.12

Na política, as pessoas às vezes sentem-se mais de "esquerda" ou de "direita". Onde você se colocaria nesta escala, onde 1 significa a esquerda e 7 significa a direita?

Esquerda Direita

1 2 3 4 5 6 7

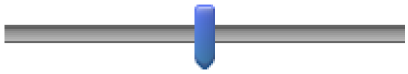
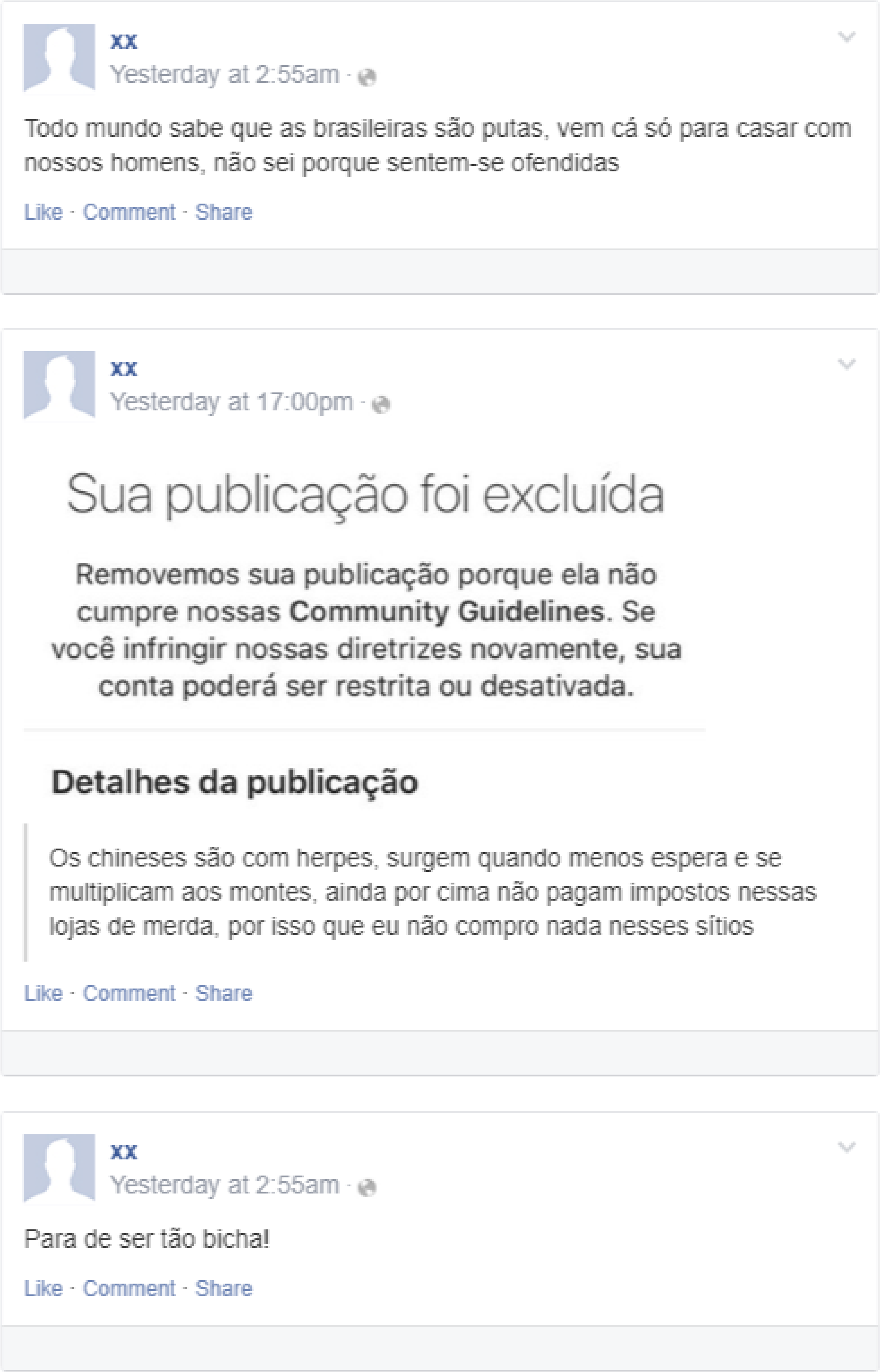
	

Figure 1

Examples of Hate Speech Posts from the Hate Speech Scale



Todo mundo sabe que as brasileiras são putas, vem cá só para casar com nossos homens, não sei porque sentem-se ofendidas

Sua publicação foi excluída

Removemos sua publicação porque ela não cumpre nossas **Community Guidelines**. Se você infringir nossas diretrizes novamente, sua conta poderá ser restrita ou desativada.

Detalhes da publicação

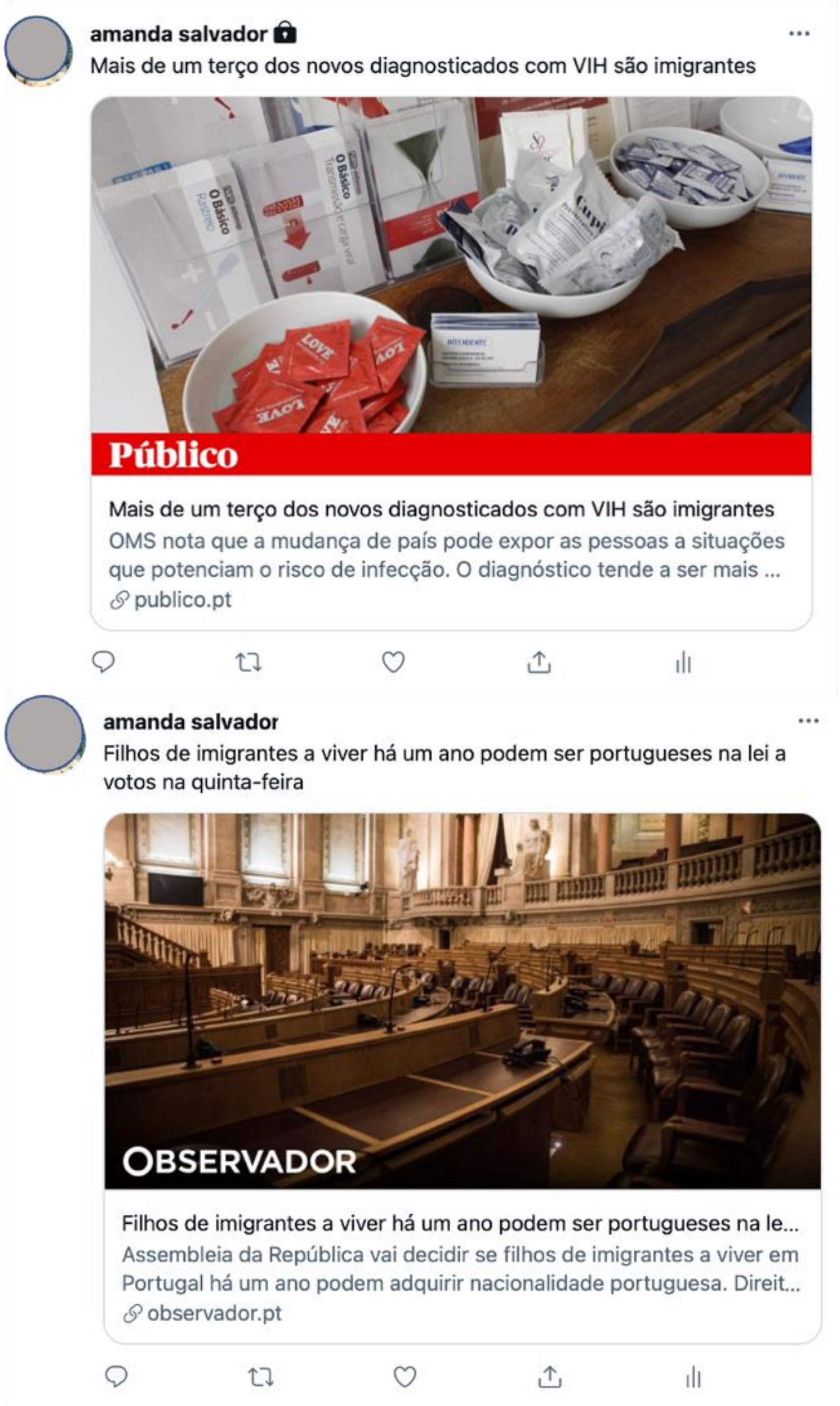
> Os chineses são com herpes, surgem quando menos espera e se multiplicam aos montes, ainda por cima não pagam impostos nessas lojas de merda, por isso que eu não compro nada nesses sítios

Para de ser tão bicha!

Note. Hate speech Facebook posts were generated with Facebook Post Simulator (https://fakedetail.com/fake-facebook-post-generator) and the post varied in two types: censored or non-censored warning, for a total of 32 unique posts. They did not have an image in the avatar and the name was standard (“xx”) to avoid bias on gender and race. Also, there was no manipulation of the number of like, share or comments.

Figure 2

Immigrant Threat Condition Manipulation Newspaper Post in a Twitter Layout



amanda salvador

Mais de um terço dos novos diagnosticados com VIH são imigrantes

Público

Mais de um terço dos novos diagnosticados com VIH são imigrantes

OMS nota que a mudança de país pode expor as pessoas a situações que potenciam o risco de infecção. O diagnóstico tende a ser mais ...

publico.pt

amanda salvador

Filhos de imigrantes a viver há um ano podem ser portugueses na lei a votos na quinta-feira

OBSERVADOR

Filhos de imigrantes a viver há um ano podem ser portugueses na le...

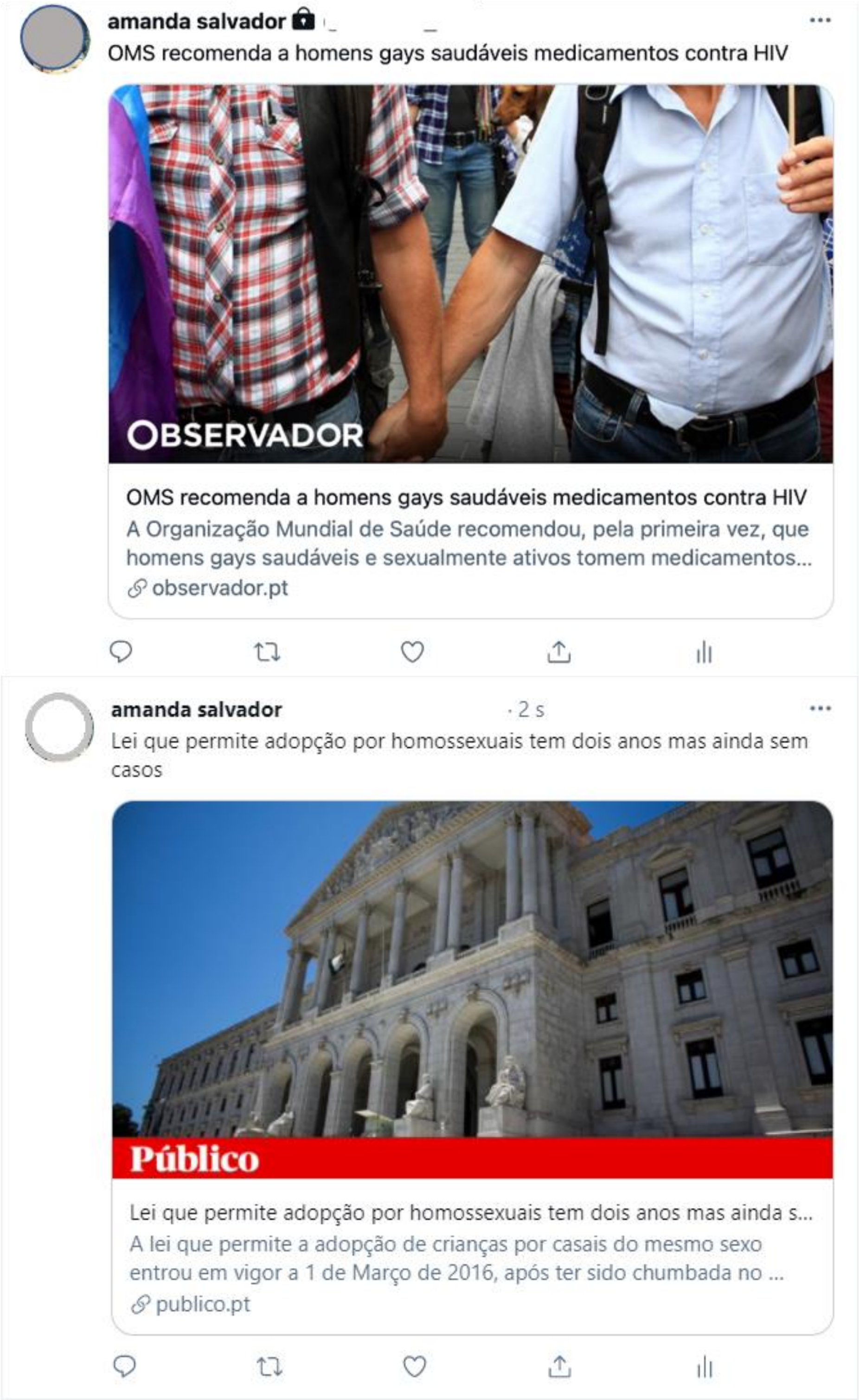
Assembleia da República vai decidir se filhos de imigrantes a viver em Portugal há um ano podem adquirir nacionalidade portuguesa. Direit...

observador.pt

Note. Stimuli were generated on the Twitter platform with the author's account and using the link of the newspaper selected for manipulation. (https://twitter.com). In the public domain

Figure 3

LGBT Rights Condition Manipulation Newspaper Post in a Twitter Layout



amanda salvador

OMS recomenda a homens gays saudáveis medicamentos contra HIV

OMS recomenda a homens gays saudáveis medicamentos contra HIV

A Organização Mundial de Saúde recomendou, pela primeira vez, que homens gays saudáveis e sexualmente ativos tomem medicamentos...

observador.pt

amanda salvador · 2 s

Lei que permite adopção por homossexuais tem dois anos mas ainda sem casos

Lei que permite adopção por homossexuais tem dois anos mas ainda s...

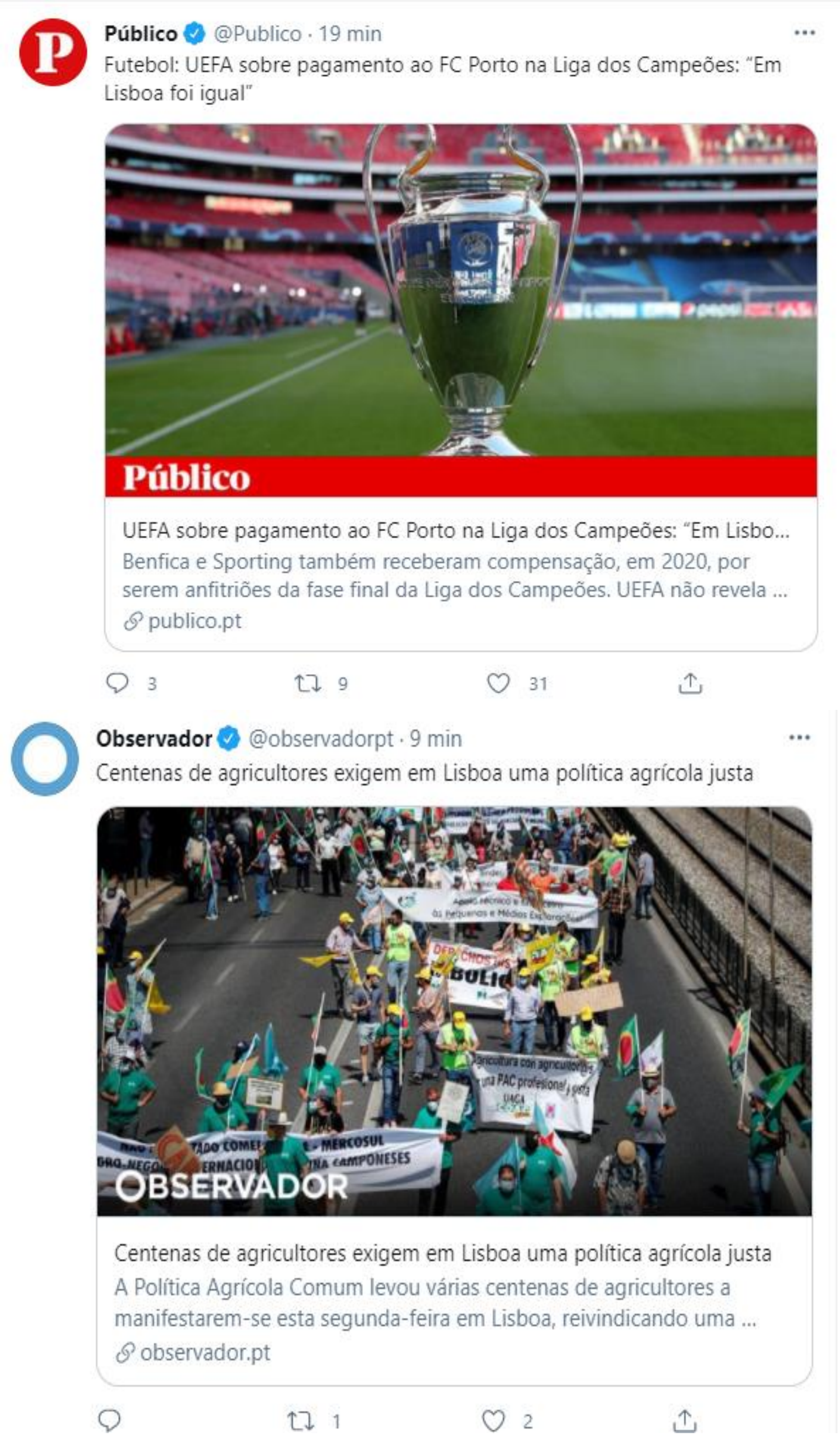
A lei que permite a adopção de crianças por casais do mesmo sexo entrou em vigor a 1 de Março de 2016, após ter sido chumbada no ...

publico.pt

Note. Stimuli were generated on the Twitter platform with the author's account and using the link of the newspaper selected for manipulation. (https://twitter.com). In the public domain

Figure 4

Control Condition Manipulation Newspaper Post in a Twitter Layout



Público @Publico · 19 min

Futebol: UEFA sobre pagamento ao FC Porto na Liga dos Campeões: "Em Lisboa foi igual"

UEFA sobre pagamento ao FC Porto na Liga dos Campeões: "Em Lisbo...

Benfica e Sporting também receberam compensação, em 2020, por serem anfitriões da fase final da Liga dos Campeões. UEFA não revela ...

publico.pt

Observador @observadorpt · 9 min

Centenas de agricultores exigem em Lisboa uma política agrícola justa

Centenas de agricultores exigem em Lisboa uma política agrícola justa

A Política Agrícola Comum levou várias centenas de agricultores a manifestarem-se esta segunda-feira em Lisboa, reivindicando uma ...

observador.pt

Note. Stimuli were generated on the Twitter platform with the author's account and using the link of the newspaper selected for manipulation. (https://twitter.com). In the public domain