U.PORTO

MESTRADO EM PSICOLOGIA
PSICOLOGIA DAS ORGANIZAÇÕES, SOCIAL E DO TRABALHO

Perspective Taking and Salience of Prescriptive Norms on Decreasing Bystander Behaviour Facing Online Hate Speech – A Lab Experiment

Beatriz Teixeira Barroso

M

2023

University of Porto
Faculty of Psychology and Educational Sciences

Perspective Taking and Salience of Prescriptive Norms on Decreasing Bystander Behaviour Facing Online Hate Speech – A Lab Experiment

Beatriz Teixeira Barroso

October, 2023

Dissertation presented for the Master's in Psychology, Organizational, Social, and Work Psychology, Faculty of Psychology and Education Sciences of University of Porto, supervised by Professor ***Catarina L. Carvalho*** and co-supervised by Professor ***Isabel R. Pinto*** (F.P.C.E.U.P.)

LEGAL NOTICES

The content of this dissertation reflects the perspectives and interpretations of the author in the moment of its submission. This dissertation may contain incorrections, both conceptual as methodological, that may have been identified after its submission. Any further utilization of its content must be exerted carefully.

By submitting this paper, the author declares that the same is resultant of her own labour, contains original contributions and all sources utilized are recognized, the same sources being correctly cited in the text and identified in the reference section. The author also declares that she doesn’t disclose in the present dissertation any contents which reproduction is vetoed by copyright or industrial propriety.

Acknowledgments

My aunt Fatima lived in a house with no more than 30 m2. She was born with a disability, which didn't allow her to do most jobs, and became a single mother early in her youth. She often told us stories of how they never had much, but always found a way to set "another seat at the table" or to give something to someone in bigger need. Her entire life was dedicated to taking care of children in her "shed", as she called it.

My maternal grandmother studied no further than the 4th grade in school. She and her siblings struggled throughout their childhoods, "even though they never starved", as she always reminds us, maybe in an attempt of consolation. She was 18 years old the first time she tasted meat, on her wedding day.

My grandfather lost his father when he was nine and had to quit school to work and help raise his siblings. He admits that there wasn't much food around in those days, and as my grandmother, didn't own a pair of shoes until his teens.

My paternal grandmother was the first person in her village to complete a superior degree. She was frequently encouraged, by her mother, to stop studying and work the fields - the land gave them what to eat; the books, didn't. Her ultimate dream was to become a doctor, but "back then, professors' children were born to be professors, doctors' children were born to be doctors, and peasants' children were born to be peasants."

This is for all (my) great educators who weren't blessed with an easy life, and still managed to raise their children in the most brilliant way.

To all my friends, family, colleagues and former professors who always encouraged me to "try again" and had faith in my abilities even in times I did not: your words had more impact than you could ever imagine, and I'm forever grateful for them.

The biggest "thank you" wouldn't suffice to Professor Isabel, Catarina and Mariana. The Social Psychology field is immensely richer because of you, and so am I.

At last, to Paulo, who fought bravely and without bitterness: you are dearly missed.

Abstract

Online hate speech is a complex phenomenon that has earned the attention of social scientists, politicians and policymakers, mainly due to its severity, increasing prevalence and difficulty in detecting. Citizens, including victims, often fail to report hate speech, which contributes to blur the real magnitude of these crimes, to informally legitimize and perpetuate its occurrence. The present study aimed to examine the efficacy of two important determinants – perspective-taking and prescriptive norms - in reducing bystander effect facing online hate speech, by promoting willingness to report hate speech situations. A lab-experiment was conducted using a fictional scenario in which participants were led to believe that they were competing in teams of four students in a European tournament with other European Universities' teams (this scenario was created and controlled by the researchers). All the remaining three members were bots. During competition, all participants were presented in a bogus team chat, with a hate speech message supposedly directed from a member to another believed to be an immigrant citizen (a situation created by the researchers). The chat included a report button in case of misbehaviour (e.g., hate speech). Participants (N = 226) were randomly assigned to one of two conditions of perspective-taking (stimulus vs. control) and one of two conditions of prescriptive norms (salience vs. control). As expected, participants used the report button more in the Perspective-taking stimulus x Salience of Prescriptive Norms condition (38) than in the control condition (21), showing that inducing perspective-taking and reinforcing prescriptive norms condemning hateful behaviour increases report behaviour (vs. bystander apathy) against online hate speech towards immigrants.

Keywords: Online Hate Speech; Lab-Experiment; Bystander Effect; Perspective Taking; Prescriptive-norms.

Resumo

O discurso de ódio online é um fenómeno complexo que tem ganho a atenção de cientistas sociais, políticos e órgãos legislativos, devido principalmente à sua severidade, prevalência crescente e dificuldade de deteção. Os cidadãos, inclusive as vítimas, muitas vezes não reportam o discurso de ódio, o que contribui para esfumar a magnitude real destes crimes, os legitima informalmente e, deste modo, perpetua a sua ocorrência. O presente estudo tem como objetivo examinar a eficácia de dois determinantes – tomada de perspetiva e mecanismos de controlo social – em reduzir o efeito do espectador face ao discurso de ódio online, promovendo, pelo contrário, o reporte destas situações. Foi feito um estudo laboratorial, recorrendo a um cenário fictício, no qual os participantes foram levados a acreditar que estavam a competir em equipas de 4 estudantes num torneio Europeu com outras equipas de universidades europeias (este cenário foi criado e controlado pelos investigadores). Os restantes 3 membros eram robôs. Durante a competição, os participantes podiam comunicar com o resto da suposta equipa através de um chat (fictício), e todos os participantes foram confrontados com uma mensagem de discurso de ódio dirigida de um membro da equipa para outro, imigrante (esta situação foi igualmente criada pelos investigadores). O chat incluía um botão de reporte no caso de se verificar alguma má conduta (ex., discurso de ódio). Os participantes (N = 226) foram aleatoriamente alocados a uma de duas condições de tomada de perspetiva (saliência vs. controlo) e a uma de duas condições de normas prescritivas (saliência vs. controlo). Conforme esperado, os participantes usaram mais o botão de reporte na condição do estímulo de Tomada de Perspetiva x Saliências de Normas Prescritivas (38) do que na condição controlo (21), demonstrando que induzir tomada de perspetiva e reforçar normas prescritivas que condenam ações odiosas aumenta o comportamento de reporte (vs. efeito do espectador) face ao discurso online dirigido a emigrantes.

Palavras-chave: Discurso de Ódio Online; Estudo Laboratorial; Efeito do Espectador; Tomada de Perspetiva; Normas Prescritivas.

Résumé

Le discours de haine en ligne est un phénomène complexe qui retient l'attention des spécialistes des sciences sociales, des politiques et des organes législatifs, principalement en raison de sa gravité, de sa prévalence croissante et de sa difficulté de détection. Les citoyens, y compris les victimes, souvent ne signalent pas les discours de haine, ce qui contribue à masquer l'ampleur réelle de ces crimes, les légitime de manière informelle et, de cette manière, perpétue leur répétition. La présente étude vise à examiner l'efficacité de deux déterminants – la prise de perspective et les mécanismes de contrôle social – pour réduire l'effet spectateur envers les discours de haine en ligne, favorisant au contraire le signalement de ces situations. Une étude en laboratoire a été réalisée, à partir d'un scénario fictif, dans lequel les participants étaient amenés à croire qu'ils concouraient en équipes de 4 étudiants dans un tournoi européen avec d'autres équipes d'universités européennes (ce scénario a été créé et contrôlé par les chercheurs). Les 3 membres restants étaient des robots. Pendant la compétition, les participants pouvaient communiquer avec le reste de l'équipe supposée à travers un chat (fictif), et tous les participants étaient confrontés à un message de discours de haine adressé d'un membre de l'équipe à un autre, un immigrant (cette situation a également été créée par le des chercheurs). Le chat comprenait un bouton de signalement en cas de mauvaise conduite (par exemple discours de haine). Les participants (N = 226) ont été assignés au hasard à l'une des deux conditions de prise de perspective (saillance vs contrôle) et à l'une des deux conditions de contrôle social (saillance vs contrôle). Comme prévu, les participants ont davantage utilisé le bouton de rapport dans la condition de stimulus Prise de perspective x Salience des normes prescriptives (38) que dans la condition de contrôle (21), démontrant qu'induire une prise de perspective et renforcer des normes prescriptives qui condamnent les actions haineuses augmente le comportement de signalement (vs. effet spectateur) face aux discours en ligne dirigés contre les émigrés.

Mots-clés: Discours de haine en ligne; Étude en laboratoire; Effet spectateur; Prise de perspective; Normes Prescriptives.

"Rooted in our tradition, some of us felt that to be abandoned by humanity then was not the ultimate. We felt that to be abandoned by God was worse than to be punished by Him. Better an unjust God than an indifferent one. For us to be ignored by God was a harsher punishment than to be a victim of His anger. Man can live far from God - not outside God. God is wherever we are. Even in suffering? Even in suffering.

In a way, to be indifferent to that suffering is what makes the human being inhuman. Indifference, after all, is more dangerous than anger and hatred. Anger can at times be creative. One writes a great poem, a great symphony. One does something special for the sake of humanity because one is angry at the injustice that one witnesses. But indifference is never creative. Even hatred at times may elicit a response. You fight it. You denounce it. You disarm it.

Indifference elicits no response. Indifference is not a response. Indifference is not a beginning; it is an end. And, therefore, indifference is always the friend of the enemy, for it benefits the aggressor - never his victim, whose pain is magnified when he or she feels forgotten. The political prisoner in his cell, the hungry children, the homeless refugees - not to respond to their plight, not to relieve their solitude by offering them a spark of hope is to exile them from human memory. And in denying their humanity, we betray our own.

Indifference, then, is not only a sin, it is a punishment."

(Elie Wiesel, 1999)

The Perils of Indifference*, *delivered 12 April 1999,
White House, Washington, D.C.

Index

Introduction

Online hate speech is currently considered one of the major issues plaguing the online space (Mathew et al., 2019), representing a dangerous and corrosive force that threatens social cohesion and individual well-being (e.g., European Comission, 2021).

Indeed, online hate speech, characterized by the use of abusive expressions that incite violence, hatred, or discrimination of people on the basis of their belonging to a social group (Kunst et al., 2021), constitutes a severe threat to the users of online spaces, causing severe psychological damage to those who are attacked (Boeckmann, 2002; Delgado, 2019, as cited in Kunst et al., 2021), traumatizing targets, evoking stress and depression (Obermaier et al., 2023). It is postulated as an intersection of several tensions: it configurates the expression of conflicts between different groups; it is a blatant example of how technologies, as the internet, have both an incredible potential, as they have a "dark side"; and it implies complex balancing between principles such as freedom of speech and the defence of the human dignity (UNESCO, 2015).

Although social media platforms, such as Facebook, use AI models to automatically detect harmful content and try to regularly update and improve these tools (Modha et al., 2020), they face several challenges (e.g., the use of slang or sarcasm - which may vary depending on language and culture -, emojis, memes, intentionally misspelled words, graphic content, diversity of targets of hate) making it impossible to detect all of the harmful content. Other measures have been implemented to address hate speech, such as employing professional content moderators and adding report buttons to their platforms, relying on users to intervene when they are exposed to abusive language (Gillespie, 2018). Thus, common users play an important role in protecting the public discourse and reducing the hateful content that is shared online (Munger, 2017). Since the existing mechanisms applied by social media platforms, such as automatic detection, seem inefficient or at least insufficient, bystander intervention becomes absolutely essential. However, citizens, including victims, often fail to report hate speech, adopting passive bystander behaviours (i.e., bystander effect), which contributes to blur the real magnitude of these crimes, and to informally legitimize and perpetuate its occurrence. Thus, it is vital to study not only the motives behind the perpetrators' misbehaviours, and how to counter it, but also the role of bystanders, who witness hate speech passively, and inadvertently contribute to the normalization and reinforcement of hateful narratives.

1. The Bystander Effect

Bystander behaviour has been studied by social psychological researchers for decades, ignited by public outrage over the 1964 murder of Kitty Genovese, that was witnessed by several of bystanders who failed to help her. Following this event, research conducted to study this phenomenon suggested that bystanders fail to act because they were either not sure if there was a real emergency, a phenomenon known as pluralistic ignorance, or because they expected other people to help the victim instead of them, a phenomenon known as diffusion of responsibility (Nickerson et al., 2014). Thus, bystander effect refers to the "inhibiting effects of the presence of others on helping behaviour" especially during ambiguous situations, leading people to look to others for social cues on how to act (Dovidio et al., as cited in Anker & Feeley, 2011).

1.1. Bystander Effect in the Context of Online Hate Speech

The bystander effect, when applied to online hate speech, refers to individuals' tendency to not report or intervene facing online hate speech.

Traditionally, research on bystander effect focuses on the explanation of people's failure to engage in prosocial behaviour towards victims in emergency situations (Latane & Darley, 1968). However, more recently, research has shifted focus to situations involving crime, namely, in the online contexts, such as cyberbullying and online hate speech (Obermaier et al., 2023; You & Lee, 2019), and how this phenomenon intensifies as the number of spectators present increases (Guazzini et al., 2019). Efforts have been made to understand if the characteristics of these new environments, such as partial or total anonymity, physical isolation, and reduced identification, may contribute to this phenomenon (Markey, 2000). For instance, previous evidence showed that people's passive behaviour decreases when they recognize crimes as real emergencies (Fischer et al., 2011) and increases when they believe it is not their duty (i.e., personal responsibility) to report crime they do not feel involved in (Siapera et al., 2018). Indeed, becoming aware of the situation and its severity, along with feelings of personal responsibility to act, are the primary requirements identified by Latané and Darley (1969) for bystanders to take action.

1.2. The Bystander Intervention Model

According to the Latané and Darley 1970's bystander intervention model, five sequential steps are needed so witnesses of misbehaviour actually take action: 1) notice the

event (i.e., to become aware that something is happening), 2) interpret the event as an emergency or a serious situation that requires intervention, 3) accept individual responsibility for intervening, 4) know and decide how to intervene or provide help, and, at finally, 5) implement intervention decisions (Latané & Darley, 1970). If any of these steps are not achieved and completed, bystanders are less likely to intervene.

In the present study, we were interested in observing the process that leads to a specific response facing online hate speech: reporting behaviour. Specifically, we examined the effect of two important determinants of prosocial behaviour in actually reporting online hate speech.

2. Determinants of Prosocial Behaviour – Reporting Online Hate Speech

Previous research (Galinsky et al., 2008; Pinto et al., 2016a,b) suggest that two determinants can be potentially effective in decreasing bystander apathy regarding online hate speech and promoting reporting behavior – perspective-taking and ingroup prescriptive norms.

Indeed, perspective-taking has been identified as one of the most important determinants for inducing prosocial behavior by increasing connection and empathy towards others (Galinsky et al., 2008). In turn, prescriptive norms determine which behaviours are tolerated or not by society and how individuals should behave in specific situations (Pinto, 2006). Therefore, highlighting hate speech as a crime, is essential so that individuals do not tolerate it and feel the need to punish offenders.

2.1. Perspective-taking

Perspective-taking is characterized by the ability to view a situation from another's point-of-view (e.g., outgroups) which is expected to impact on several skills as empathy, sympathy and caring, bias and stereotypes reduction, fostering more favourable attitudes (e.g., towards outgroups) (Galinsky et al., 2008), and on anti-discrimination norms and policies (Abrams et al., 2014) that proscribe tolerance for discriminatory behaviour. Thus, this capacity to put oneself in “another person’s shoes”, plays an important role in effective social interactions, by leading people to alter their behavior in social situations as the result of acknowledging the perspective of the other, being also considered as one fundamental source of empathy (Batson, 1991, as cited in Malle & Hodges, 2007). Empathy, in its turn, is considered a vital predecessor to prosocial behavior (Eisenberg & Miller, 1987). Therefore, by stimulating perspective-taking, we expect to lead participants to engage more in prosocial behaviors such as reporting an online hate speech situation.

2.2. Prescriptive Norms

Prescriptive norms are defined as propositions that regulate and control interpersonal relationships within similar contexts (Pinto, 2006). The main functions of norms include the definition of which behaviour is tolerated or considered as desirable, the reduction of possible conflicts or misunderstandings, and diminishing the costs of unnecessary communication (Pinto, 2006). Once there is established a norm, and group members internalize it, there are no longer controversies regarding to behaviours associated with this norm (Pinto, 2006). However, the existence of effective social control mechanisms (i.e., punishing norms violations) are essential to maintain and reinforce individuals' commitment to prescriptive norms (e.g., hate speech is a crime) (Pinto et al., 2016a). Social control mechanisms implemented by social media platforms are very often perceived as ineffective, inexistent, or as not punitively relevant, leading people to hesitate to report misbehaviour (Siapera et al., 2018). Thus, highlighting the prescriptive norm and which behaviour is most appropriate (i.e., not to tolerate hate speech) seems to be essential to counter bystanders' apathy (i.e., bystander effect) and increase motivation to report and punish offenders.

3. The present study

This work is part of a major project ""VIGILANT CITIZENS AGAINST HATE": How to counter bystander apathy and increase citizens' commitment against online hate speech?", funded by "la Caixa" Foundation and the Portuguese Foundation for Science and Technology (FCT) (SR20 - Social Research 2020; ref. SR20-00136). The project aimed to test the effectiveness of some potential determinants of prosocial behaviour in stimulating bystanders' moral self-regulation, leading them to act against online hate speech.

In this study, we test the impact of salience of perspective-taking and of prescriptive norms on increasing moral engagement towards reporting witnessed online hate speech (i.e., prosocial behaviour).

We expect that perspective-taking regarding a member of a socially vulnerable or minority group – potential targets of hate speech – and the salience of prescriptive ingroup norms about how members are expected to proceed facing hate speech, should enhance intolerance regarding misbehaviour and consequently fell more motivated to report them. In

order to test the preceding idea, we conducted a laboratory experiment, in which we intended to replicate an online hate speech situation, similar to those that occur in a natural environment. In this sense, we created a scenario of a fictitious team game.

The actual reporting behaviour of online hate speech was measured, as well as a self-reporting measure about the same behaviour. We decided to include both measures given the importance of studying real behaviour, compared to self-report behaviour, in social sciences.

Based on existing literature (Pinto, Marques, & Paez, 2016; Galinsky et al, 2008), we expect that bystanders' apathy (i.e., bystander effect), when faced with online hate speech, should decrease when empathy is stimulated (i.e., perspective-taking) and prescriptive norms are salient. Specifically, it is expected that the participants under these conditions adopt a more prosocial behaviour (report hate speech more) than the participants in the control condition (empathy is not stimulated and prescriptive norms are not salient).

Taking into account the Bystander Intervention Model (Latané & Darley, 1970), we expect each step of the model to be associated with the preceding steps, and, more importantly, that the reporting behavior to be related to Step 4 of the model (know and decide how to intervene).

Method

1. Participants and Design

Participants were 251 females and 34 males (8 that identified as "other") first- and second-year students (N = 293) enrolled in a Psychology course (convenience sample), aged from 18 to 58 years-old (M = 20.35, SD = 5.08). Since the experiment involved a hate speech situation against immigrants, all participants who initially indicated another nationality did not participate in the main study and completed another non-related task. Only national participants completed the main study.

Then, participants were randomly assigned to one of the two of perspective-taking (stimulus vs. control) and one of the two conditions of prescriptive norms (salience vs. control). We discarded one participant who failed both manipulations-checks after the Perspective-taking stimulus was introduced. Thus, remained in the analysis, 226 Portuguese nationals (191 females, 27 males), aged from 18 to 58 years-old (M = 20.30, SD = 5.54). Participants by

condition ranged between *n* =51 (Control x Control condition) and *n* = 62 (Perspective taking stimulus x Control condition).

Regarding their political tendencies, participants positioned as standing more as left-wing (*M* = 3.19; *SD* = 1.26; 1 = *More to the left,* 7 = *More to the* r*ight*). Participants were also asked about perceived socioeconomic status compared to other citizens of the country where they live on a 7-point Likert-scale (1 = *Very low*, 7 = *Very high;* around the scale midpoint, *M* = 4.15, *SD* = .90).

Participants' sex, age, political orientation and socioeconomic status did not significantly differ across conditions, $\chi 2(6) = 1.71$, $p = .944$, $F(3, 219) = 1.80$, $p = .149$, $F(3, 223) = 0.14$, $p = .936$; $F(3, 222) = 0.19$, $p = .900$, respectively.

2. Procedures

Participants were recruited to participate in the study as a curricular activity. Participants came to an experimental room, in groups up to 15, specifically designed for the study and equipped with computers. Upon their arrival, they were told that they would participate in two unrelated tasks. In fact, both tasks were part of the main study. Participants were placed in a computer previously prepared for the effect. After all the procedure and instructions had been explained, participants were asked to sign the informed consent (which can be consulted in Appendix 1).

2.1. Task 1 and Perspective-taking Manipulation

Participants were informed that their first task would be an opinion study about socially relevant issues. After completed demographic information, participants were included in one of two Perspective-taking conditions. Half of the participants were presented with a text based in *Alien Simulation* from Hillman and Martin (2002) and Hodson and colleagues (2009)'s work that combine perspective-taking techniques and role-playing exercises (stimulus condition). Specifically, participants were asked to imagine life in an alien planet named Dorothy, in which 3000 humans landed and had to forcefully live, encountering a living species very similar to them, named Dorys. In this planet, humans were victims of discrimination and hate speech, having to survive in similar conditions to those faced by migrants in real life. The purpose of this story was to stimulate empathy towards those who experience similar situational

constraints (e.g., discrimination and hate speech) such as immigrants in a foreign country (see the full text in Appendix 2). Then participants answered to measures unrelated to the main study just to add realism to the task (e.g., social dominance orientation, belief in a just world scales). The other half of the participants only responded to the measures (control condition).

2.2. Task 2 and Salience of Prescriptive Norms Manipulation

The second task was presented as a study aimed at examining the idea that conducting cognitive stimulation activities as a member of a team (opposed to developing the same activities individually) enhances individuals' productivity, and personal and social well-being. Participants were led to believe that, to test this idea, they would participate in the University Games, part of the European Universities Olympics (EUO) project. Participants were led to believe that they would participate in randomly generated teams of 4 students from the same University, that were available to play in the same schedule as them.

It was explained that these games consisted in a European competition that occurs every two years, which aims to test the levels of general knowledge of the university students. In the current edition, 48 universities throughout Europe were participating, including 5 Portuguese Universities (University of Minho, University of Trás-os-Montes and Alto Douro, University of Coimbra, University of Porto, and University of Lisboa). Participants were told that the Games would be in a Quiz format (multiple choice questions). The Quiz was divided into 3 rounds, with 10 questions each, about several themes of general knowledge. Participants had 2 minutes to answer the questions in each round. Only the correct answers would be considered, and each would be rewarded with 5 points. Individual scores would never be revealed; instead, group scores would be shown at the end of each round that corresponded to the sum of the scores of all team members. The complete initial instructions can be found at Appendix 3.

2.2.1. Team chat

Participants were led to believe that the Games platform included a chat where members of the same team would have the opportunity to talk to each other during the competition. Before the beginning of the Games and at the end of each round, the team could communicate through this chat, that was private and anonymous. All chat interactions (conversations), apart from the participant him/herself, were created by the researchers – that is, all the other

supposedly team members were bots. The chat was created in Qualtrics using JavaScript and HTML language and dozens of combinations were prepared so that what appeared to each participant on the chat page (i.e., landing page for conversation) was automatically adapted to the participant's behavior (e.g., whether s/he wrote in the chat and interacted with the team members or just read what others wrote). Team members' identification was hidden and each one was identified only with a number (e.g., Participant 1, Participant 2, …). All participants were informed that they were Participant 2.

2.2.2. Salience of prescriptive norms

After participants read the instructions about the Team Chat, and before starting their participation in the Game, participants read further information. Namely, participants read that the team chat platform included one button to contact directly, and privately, any other team member (message button); and another button that would allow anonymously reporting any irregularity or misconduct by any member in the team chat (report button). Then, in the Salience of Prescriptive Norms condition, participants also read that "The University is governed by an ethical code of academic conduct, and all its members are expected to maintain an inclusive stance, rejecting and sanctioning any discriminatory practices, harassment, intimidation, retaliation, physical violence, or moral coercion. Offensive or discriminatory language that incites violence or hatred and threatens the well-being of participants should not be tolerated. Any violation of the code of conduct must be reported by the participants.". The participants in the control condition did not have access to this information.

2.2.3. Hate Speech Situation

During the chat interactions, participants were led to believe that one of the team members was an immigrant (Brazilian citizen; always identified as Participant 3). This specific member interacted, in the team chat, using Brazilian-Portuguese language and typical cultural expressions, so that nationality was noticeable and clear for the participants. In the final moment the Team Chat was available, following the final round of the Games, a hate speech situation was presented against the immigrant participant, perpetrated by a supposed Participant 1 (the offender). The final score of the Games had just been revealed, and Participant 1 rebelled against the immigrant team member, blaming her/him for the score, writing in the chat: "153 points out of 200? If it weren't for the Brazilian, we would have more!

And these stupid Brazilians are stealing places at University of Porto". A few seconds later, participants received the information that the team chat would close in 10 seconds. Thus, they had to decide whether to take action - use the report button, publicly condemn the offender in chat - or remain passive bystanders.

After the team chat closed, participants answered a set of questions aimed at accessing the different steps of the bystander intervention model, adapted to the context of our study.

Debriefing and explanation of the aim of the research took place after data collection ended to prevent the experimental procedure from being revealed.

3. Measures

3.1. Control Measures

3.1.1. Perspective-taking Manipulation Check

To ensure that participants, in the perspective-taking stimulus condition, had read and understood the text, we included two manipulations-checks by asking participants to indicate 1) Why was a colleague of yours kicked off a bus on planet Dorothy? (response options: 1 = *For praying out loud*.; 2 = *For speaking Portuguese with his wife*.; 3 = *For claiming for better living conditions*.); 2) What did the Mayor say about your complaints? (1 = *He agreed with us and promised better working conditions*.; 2 = *He kicked us off the planet*.; 3 = *He claimed that we must make a collective effort to fully integrate ourselves*.). Only those participants who failed both manipulation checks were excluded.

3.1.2. Prescriptive Norms Manipulation Checks

To ensure that participants had read all the instructions and the information about the prescriptive norms (Salience of Prescriptive Norms condition) we included three questions presented to all participants (e.g., "How much time do you have to answer each round of questions?") that would allow to not raise any suspicion regarding this manipulation and one question to remember the existence and functions of the chat buttons ("What are the functions of the buttons in the team chat?"), presented only to participants in the Salience of Prescriptive Norms condition. If the participants answered any of these questions incorrectly, they were presented again with a summary of the instructions.

3.2. Dependent Measures

3.2.1. Report Button (Actual Behavior)

The team chat included a button to report misconduct that allowed us to measure the actual behaviour of the participants. We categorized responses according to whether or not they used the button (1 = *Did not use the button*; 2 = *Used the button*).

When pressing this button, participants were also presented with several options of misconduct (multiple choice): violent language, sexual harassment, moral harassment, incitement to suicide or self-harm, false information, hate speech, intimidation, threat and other (in which participants could write what they thought that could be best fitted).

3.2.2. Notice Misconduct

To maintain the psychological realism of our cover story, we asked participants' opinion about the Games (e.g., "Regarding the time available to answer each round of questions, please indicate to what extent this time was adequate.") and about the Team Chat (e.g., "Do you think that the contact and communication with the rest of your team increased your motivation during the Games?"). These measures were not considered for the analyses.

Then, participants were asked whether they had noticed any irregularities or misconduct during the team chat (1 = *Yes*; 2 = *No*; 3 = *Not sure*).

Next, participants read the information that at least one person of their team had reported misconduct, identified as a hate speech situation in the Team Chat. This information would allow to justify the need for their responses regarding the hate speech situation.

3.2.3. Bystander Intervention Model (Self-reported Measures)

Participants answered a set of questions designed to assess each step of the Bystander Intervention Model (Latané & Darley, 1970) adapted to this specific context.

We assess Step 1 of the Bystander Intervention Model (Notice and identify hate content online) through 3 items: 1) "Did you notice any situation, during the team chat, that may be interpreted as hate speech?", 2)"Are you sure that this situation was, in fact, an online hate speech occurrence?", 3) "Was it easy to identify this situation as an online hate speech occurrence?".

Step 2 (Interpret hate speech as a serious occurrence) was measured through 4 items: 4) "In the case of having observed a hate speech occurrence, do you think that it was severe?" 5) In the case of having observed a hate speech occurrence, do you believe that such situation

had a negative impact in the participant that was the victim of it?" 6) In the case of having observed a hate speech occurrence, do you believe that such situation may hurt its victim?" 7) In the case of having observed the hate speech situation, do you believe that it's more likely that it was just a joke?".

Step 3 (Accept responsibility to help) was measured through 4 items: 8) "In case you have observed the hate speech occurrence, do you believe it's your responsibility to intervene?" 9) "Even if you weren't the author of the hate speech occurrence, do you believe that it's also your responsibility to intervene to stop the situation?" 10) "In the case of having observed the hate speech situation, did you felt that you should do something to stop it?" 11) "Even if you weren't the author of the hate speech occurrence, do you believe it was your duty to stop it?".

Step 4 (Know how to help) was measured through 3 items: 12) "Faced with the situation of hate speech, did you know what to do?"; 13) "Faced with the situation of hate speech, did you know how to report/denounce it?"; 14) "Faced with the situation of hate speech, did you know how to intervene to stop the situation?". All the 14 items were answered on a 7-point Likert-scale ranging from 1 (*No, not at all*) to 7 (*Yes, for sure*).

A principal components factorial analysis conducted on these items extracted three factors accounting for 69% of the total variance. STEP 1 and STEP 2 loaded together on the same factor corresponding to Factor 1 (32% of variance); STEP 3 corresponded to Factor 2 (21%); and STEP 4 corresponded to Factor 3 (16%). Although the items corresponding to Step 1 and 2 of the Bystander Intervention Model loaded together, we created two independent variables, one for each step, following the theoretical structure proposed by the Bystander Intervention Model (Latané & Darley, 1970).

Finally, in order to assess Step 5 (Implement intervention decision – self-reported measure) participants were asked to indicate whether they had reported any misconduct (i.e., if they had used the report button) (1 = *No*, 2 = *Yes*). Those who answered "Yes" were also asked to indicate what kind of misconduct they witnessed by selecting one of the options (e.g., hate speech).

Results

1. Participants' Actual Behaviour

A Chi-square test showed that the proportion of participants who used the report button differ by experimental condition, X^2 (3, N = 226) = 7.95, p = .047. In order to verify precisely which percentages differ significantly from each other, we performed pairwise Z-Tests using Bonferroni correction. As we can see in Table 1, the analysis revealed that the percentage of participants who used the report button in the Perspective-taking stimulus x Salience of Prescriptive Norms condition is significantly higher (68% from n = 38) for those who used the report button in the Control x Control condition (41% from n = 21), p = .033.

Table 1

Number of Participants Who Used the Report button by Experimental Condition

	Experimental condition				
Used the Report button (actual behavior)	COND 1	COND 2	COND 3	COND 4	Total
Did not use the button	**30** (59%)	29 (47%)	**18** (32%)	24 (42%)	101
Used the button	**21** (41%)	33 (53%)	**38** (68%)	33 (58%)	125
Total	51	62	56	57	226

Note: COND 1 = *Control x control condition*; COND 2 = *Perspective-taking stimulus x control condition*; COND 3 = *Perspective-taking stimulus x Salience of prescriptive norms condition*; COND 4 = *Control x Salience of prescriptive norms condition*.

We decided to further inspect the effect of each factor to best assess the impact of each determinant on participants' actual behaviour. Analysing the use of the report button between the Perspective-taking conditions, results showed that the proportion of participants who used the report button did not differ within the Perspective-taking conditions (Stimulus vs Control), X^2 (3, N = 226) = 2.36, p = .125 (see Table 2).

Table 2

Number of Participants Who Used the Report Button: Perspective Taking vs Control

	Perspective-taking Condition		
Used the Report button (actual behavior)	Stimulus	Total	Total
Did not use the button	54 (50%)	47 (40%)	101
Used the button	54 (50%)	71 (60%)	125
Total	108	118	226

Regarding the Salience of Prescriptive Norms factor, a Chi-square test showed that the proportion of participants who used the report button differ within the Salience of Prescriptive Norms condition (Stimulus vs Control), X^2 (3, N = 226) = 7.95, p = .047. In order to verify precisely which percentages differ significantly from each other, we performed pairwise Z-Tests using Bonferroni correction. As we can see in Table 3, the analysis revealed that the percentage of participants who used the report button in the Salience of Prescriptive Norms condition is significantly lower (37% from n = 42) for those who used the report button in the Control x Control condition (52% from n = 59), p = .023.

Table 3

Number of Participants Who Used the Report Button: Salience of Prescriptive Norms condition vs Control condition

	Prescriptive Norms Condition		
Used the Report button (actual behavior)	Control	Salience of prescriptive norms	Total
Did not use the button	59 (52%)	42 (37%)	101
Used the button	**54** (48%)	**71** (63%)	125
Total	113	113	226

Finally, when using the report button, participants could also indicate the type of misbehaviour observed by selecting more than one option. They could choose more than one

option. Of those who used it, as expected, the majority (74%) reported hate speech, 11% reported intimidation, 6% moral harassment, 4% spread of false information, 3% threats, and 3% reported other misbehaviour (e.g., xenophobia, inappropriate language). Thus, participants correctly identified and report the situation presented as hate speech.

2. Notice Misconduct

When asked whether they had observed any misconduct during chat interactions, 90% of the participants indicated "Yes", only 4% indicated "No", and 7% stated they were not entirely sure. The Chi-Square test showed no statistically significant differences between experimental conditions regarding notice misconduct (which corresponds to the first step of the Bystander Intervention Model – notice the event), X^2 (6, N = 224) = 5.89, p = .436.

3. Bystander Intervention Model Self-Reported Measures

A one-way between subjects ANOVA was conducted to evaluate the effect of the experimental conditions on each step of the Bystander Intervention Model. The means and standard deviations are presented in Table 4 below. Results showed only a significant effect of the experimental conditions on STEP 2, $F(3, 218) = 4.35$, $p = .005$. Post hoc comparisons using Tukey test indicated that the mean score of STEP 2 for the Control x Salience of prescriptive norms condition ($M = 6.00$, $SD = 1.14$) was significantly lower than the remaining experimental conditions, namely, the Control x Control ($M = 6.46$, $SD = 0.73$, $p = .034$), Perspective-taking stimulus x Control ($M = 6.41$, $SD = 0.87$, $p = .053$), and the Perspective-taking stimulus x Salience of prescriptive norms ($M = 6.54$, $SD = 0.56$, $p = .006$) conditions.

Table 4

Summary of Means and Standard Deviations of Bystander Intervention Model Steps by Experimental Condition

	Experimental condition							
	COND 1		COND 2		COND 3		COND 4	
	n = 50		n = 61		n = 54		n = 57	
Variable	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***

STEP 1	6.51	0.84	6.53	0.93	6.55	0.81	6.16	1.24
STEP 2	6.46	0.73	6.41	0.87	6.54	0.56	**6.00**	1.14
STEP 3	6.56	0.78	6.34	0.91	6.44	0.81	6.23	1.05
STEP 4	5.30	1.23	5.41	1.36	5.70	1.10	5.57	1.11

Note 1: COND 1 = *Control x control condition*; COND 2 = *Perspective-taking stimulus x control condition*; COND 3 = *Perspective-taking stimulus x Salience of prescriptive norms condition*; COND 4 = *Control x Salience of prescriptive norms condition*.

STEP 1 (Notice and identify hate content); STEP 2 (Interpret the situation as serious); STEP 3 (Accept responsibility to intervene); STEP 4 (Know how to intervene).

Regarding STEP 5 (Implement intervention - self-reported measures of Reporting behaviour), 77% (n =175) of the total sample indicated having used the report button. A Chi-square test showed that the proportion of participants who indicated having used the report button did not differ by experimental condition, X^2 (3, N = 224) = 1.52, p = .677 (see Table 5). Interestingly, the number of participants who indicated having used the report button (175) was higher than the number of participants who actually used the report button (i.e., n = 125). This result emphasizes the importance of using measures of actual behaviour in addition to self-report measures.

Table 5

Number of Participants Who Indicated Having Used the Report Button (Self-Reported Measure) by Experimental Condition.

	Experimental condition				
Used the Report button (Self-reported measure)	COND 1	COND 2	COND 3	COND 4	Total
No	14 (6%)	14 (6%)	10 (5%)	11 (5%)	49
Yes	37 (17%)	48 (21%)	44 (20%)	46 (20%)	175
Total	51	62	56	57	224

Note: COND 1 = *Control x control condition*; COND 2 = *Perspective-taking stimulus x control condition*; COND 3 = *Perspective-taking stimulus x Salience of prescriptive norms condition*; COND 4 = *Control x Salience of prescriptive norms condition*.

By observing the product-moment correlations between all the steps of the bystander intervention model, we observed that all the steps are significantly correlated to the subsequent step of the model, across all experimental conditions, except for STEP 4 (Know how to help) that is not correlated to STEP 5 (Self-Reported measure) in the Perspective-taking stimulus x Control condition (r = .14, p = .287) and in Perspective-taking stimulus x Prescriptive Norms condition (r = .26, p = .054). Nevertheless, STEP 4 (Know how to help) is significantly correlated to the measure of actual behaviour (i.e., use of the report button) across all experimental conditions. Due to space constraints, correlations tables are presented in Appendix 4.

Discussion

The results suggest that the combination of perspective-taking stimulation and salience of prescriptive norms seems crucial for enhancing reporting behaviour, when compared to the absence of any stimulus. Indeed, those exposed to the perspective-taking stimulus and that were primed with prescriptive norms, report more the hate speech situation than the participants who were not presented with these stimuli.

Results also showed that perspective-taking stimulus in isolation may not be as effective in promoting reporting behavior as making prescriptive norms salient, which appears to be a critical determinant to potentiate reporting behaviour. However, participants in the Control x Prescriptive Norms condition interpreted the event as less severe (Step 2 of the Bystander Intervention Model) than the participants in remaining conditions. These results seem to suggest that if prescriptive norms are salient, but perspective-taking or empathy are not stimulated, participants are able to notice the event (Step 1) but may disregard its severity. Indeed, no differences were found on the measure of Notice misconducted across experimental conditions and the majority of participants (90%) indicated to have noticed the hate speech situation.

As expected, all the steps of Bystander Intervention Model were found to be correlated with the following step across experimental conditions. However, Step 4 (Know How to Help) was not correlated to the self-report measure of Step 5 (Implement Action) in the Perspective-taking stimulus x Control condition and in the Perspective-taking stimulus x Prescriptive Norms condition, although the measure of actual behavior was.

These results also emphasize the relevance of including both self-report measures and measures of actual behavior in this type of investigation. Indeed, we observed that 175 participants stated that they had reported the online hate speech occurrence, when, in fact, only 125 participants actually used the report button. We can think that this may have occurred due to social desirability bias, leading participants to answer according to researcher's expectations, rather than their own beliefs or real behaviour (i.e., the use of the report button).

Conclusions

The aim of our research was to examine the efficacy of two important determinants – perspective-taking and prescriptive norms – in reducing bystander apathy facing online hate speech, by promoting willingness to report such situations. Taking together, results seem to suggest that, as predicted, both perspective-taking stimulus and salience of prescriptive norms are important determinants of reporting behaviour, especially when used together. Thus, if perspective-taking is encouraged and prescriptive norms are salient, when faced with an online hate speech situation, individuals will be more likely to adopt prosocial behaviors such as reporting the incident.

Results also showed that participants noticed equally the online hate speech occurrence throughout the experimental conditions but interpreted it as less severe (i.e., Step 2 of the Bystander Intervention Model) when only the prescriptive norms were salient. In other words, it seems that highlighting prescriptive norms in isolation (without encouraged perspective-taking) may lead individuals to downplay the severity of the hate speech situation which in turn may affect subsequent steps, such as decrease their motivation to accept the responsibility to intervene facing such situations. These results emphasize the need to pairing both prescriptive norms and perspective-taking. Indeed, results showed that, although salience of prescriptive norms is crucial in predicting reporting behaviour, when used alone may have a perverse effect on some of the necessary stages for bystanders to intervene. However, this negative effect seems to be counteracted by perspective-taking stimulus.

1. Limitations and Directions for Future Research

Despite the potential contribution of our results, there are potential limitations that should be addressed in future research.

This study employed a convenience sample primarily composed of female students. While a convenience sample was necessary for conducting the experimental study, the sample's demographic composition restricts the generalizability of the findings to a broader population. Future research should strive for a more diverse and representative sample to enhance the external validity of the results. Additionally, the participants being psychology students might influence their responses and behaviour, potentially deviating from how individuals from other age groups and backgrounds might react. To address this, future studies could include participants from different age groups and backgrounds.

Furthermore, while efforts were made to create a scenario that closely resembled real-life situations, the credibility and authenticity of our scenario might not have been equally persuasive to all participants. Some individuals might have been sceptical about the authenticity of the experimental scenario, potentially affecting their responses and behaviours. To mitigate this limitation, future research could create a scenario that is more similar to real-life social networks.

It would also be interesting, in further studies, to test different determinants (e.g., human rights salience), since scientific literature enhances several other determinants as being able to counter bystander effect against online hate speech.

2. Concluding Remarks

We believe our results are promising to inform social media platforms and Governments about additional strategies to combat online hate speech, emphasizing the need to reinforce empathy in social media users and not only create tougher standards (prescriptive norms) but also disseminate them clearly and persistently for a safer navigation through social networks. Indeed, results can be informative and useful, for example, in sensibilization campaigns to counter online hate speech and encourage reporting behavior, making bystanders more accountable, and on urging decision-makers to consider harsher rules and penalties for

online hate speech, since currently, online hate speech is not yet considered a crime in European Union (Committee of Ministers of the Council of Europe, 2023).

References

Abrams, D., Travaglino, G. A., Randsley De Moura, G., & May, P. J. (2014). A step too far? Leader racism inhibits transgression credit. *European Journal of Social Psychology, 44*(7), 730–735. https://doi.org/10.1002/ejsp.2063

Anker, A. E., & Feeley, T. H. (2011). Are Nonparticipants in Prosocial Behavior Merely Innocent Bystanders? *Health Communication, 26*(1), 13–24. https://doi.org/10.1080/10410236.2011.527618

Assimakopoulos, S., Baider, F. H., & Millar, S. (2017). *Online Hate Speech in the European Union*. Springer International Publishing. https://doi.org/10.1007/978-3-319-72604-5

Brewster, M., & Tucker, J. M. (2016). Understanding Bystander Behavior: The Influence of and Interaction Between Bystander Characteristics and Situational Factors. *Victims & Offenders*, *11*(3), 455–481. https://doi.org/10.1080/15564886.2015.1009593

Comission, E. (2021). *COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL A more inclusive and protective Europe: Extending the list of EU crimes to hate speech and hate crime*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021DC0777&qid=1639350415533

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin, 101*(1), 91–119. https://doi.org/10.1037/0033-2909.101.1.91

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, *137*(4), 517–537. https://doi.org/10.1037/a0023304

Freis, S. D., & Gurung, R. A. R. (2013). A Facebook analysis of helping behavior in online bullying. *Psychology of Popular Media Culture*, *2*(1), 11–19. https://doi.org/10.1037/a0030239

Galinsky, A. D., Maddux, W. W., Gilin, D., & White, J. B. (2008). Why It Pays to Get Inside the Head of Your Opponent: The Differential Effects of Perspective Taking and Empathy in Negotiations. *Psychological Science*, *19*(4), 378–384. https://doi.org/10.1111/j.1467-9280.2008.02096.x

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press. https://books.google.pt/books?id=-RteDwAAQBAJ

Guazzini, A., Imbimbo, E., Stefanelli, F., & Bravi, G. (2019). The Online Bystander Effect: Evidence from a Study on Synchronous Facebook Communications. Em S. El Yacoubi, F. Bagnoli, & G. Pacini (Eds.), *Internet Science* (Vol. 11938, pp. 153–167). Springer International Publishing. https://doi.org/10.1007/978-3-030-34770-3_12

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

Latane, B., & Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology*, *10*(3), 215–221. https://doi.org/10.1037/h0026570

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* Appleton-Century Crofts.

Malle, B. F., & Hodges, S. D. (2007). *Other Minds: How Humans Bridge the Divide Between Self and Others*. Guilford Publications. https://books.google.pt/books?id=JJqB_NboEYYC

Markey, P. M. (2000). Bystander intervention in computer-mediated communication. *Computers in Human Behavior*, *16*(2), 183–188. https://doi.org/10.1016/S0747-5632(99)00056-4

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of Hate Speech in Online Social Media. *Proceedings of the 10th ACM Conference on Web Science*, 173–182. https://doi.org/10.1145/3292522.3326034

Modha, S., Majumder, P., Mandl, T., & Mandalia, C. (2020). Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance. *Expert Systems with Applications*, *161*, 113725. https://doi.org/10.1016/j.eswa.2020.113725

Munger, K. (2017). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

Nickerson, A. B., Aloe, A. M., Livingston, J. A., & Feeley, T. H. (2014). Measurement of the bystander intervention model for bullying and sexual harassment. *Journal of Adolescence*, *37*(4), 391–400. https://doi.org/10.1016/j.adolescence.2014.03.003

Obermaier, M., Schmuck, D., & Saleem, M. (2023). I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *New Media & Society*, *25*(9), 2339–2358. https://doi.org/10.1177/14614448211017527

Pinto, I. R. (2006). *A Subjective Group Dynamics Approach to Group Socialization Processes (thesis).* (Porto). University of Porto.

Pinto, I. R., Marques, J. M., Levine, J. M., & Abrams, D. (2016b). Membership role and subjective group dynamics: Impact on evaluative intragroup differentiation and commitment to prescriptive norms. *Group Processes & Intergroup Relations*, *19*(5), 570–590. https://doi.org/10.1177/1368430216638531

Pinto, I. R., Marques, J. M., & Paez, D. (2016a). National identification as a function of perceived social control: A subjective group dynamics analysis. *Group Processes & Intergroup Relations*, *19*(2), 236–256. https://doi.org/10.1177/1368430215577225

Siapera, E., Moreo, E., & Zhou, J. (2018). *Hate Track—Tracking and Monitoring Racist Speech Online. Reasearch Repository UC* (Reasearch Repository UCD).

UNESCO. (2015). *Countering online hate speech*. United Nations Educational, Scientific and Cultural Organization. https://books.google.pt/books?id=WAVgCgAAQBAJ

Wiesel, E. (1999). *American Rhetoric: Elie Wiesel—The Perils of Indifference*. https://www.americanrhetoric.com/speeches/ewieselperilsofindifference.html

You, L., & Lee, Y.-H. (2019). The bystander effect in cyberbullying on social network sites: Anonymity, group size, and intervention intentions. *Telematics and Informatics*, *45*, 101284. https://doi.org/10.1016/j.tele.2019.101284

Appendix

Appendix 1

Information about the Study and Informed Consent:

Introduction and Context: The present studies are being conducted by the Social Psychology Research Group of the Faculty of Psychology and Educational Sciences at the University of Porto (FPCEUP).

Study Objectives: The studies aim to gather the opinions of citizens on relevant social issues.

Procedures: Participation in these studies involves responding to relevant social questions. You will also be asked for some sociodemographic information such as age, gender, and nationality. We will also request information regarding education, employment status, and political orientation. At no point will you be asked for your name, email address, or any other personally identifying information.

Eligibility: Any resident citizen of Portugal who is at least 18 years old may participate in this study.

Risks and Benefits: There may be minor risks associated with participating in this study. Some participants may feel some discomfort when answering certain topics. In this regard, if you wish to terminate your participation, please be aware that you can do so at any time without any consequences. While this study may not personally benefit you, we hope that the results will contribute to a better understanding of certain psychological processes associated with important social phenomena. We also hope that participation in this study will be interesting and informative and may allow you to reflect on important day-to-day issues.

Payment or Compensation: You will not receive any material rewards for participating in these studies.

Voluntary Participation: Participation in these studies is entirely voluntary. You are free to refuse to participate or to stop responding at any time.

Confidentiality and Data Protection: The questionnaire is implemented on a platform managed by Qualtrics, subject to the license conditions subscribed by FPCEUP. Your responses will be downloaded from the platform to the computer of the responsible researchers, where they will be analyzed in an aggregated manner, i.e., together with the responses given by all individuals participating in the study. Each participant will only be identified by an alphanumeric code assigned automatically and randomly by the questionnaire platform. Furthermore, the responsible researchers commit to treating all information confidentially. The data will be stored and retained only for the period necessary to fulfill the purposes for which they were collected and processed or for up to five years after data collection is completed.

Purpose of Data Processing and Dissemination of Results: Data collection and processing are carried out exclusively for scientific research purposes. The final results of the study may be published in scientific journals and academic newspapers, presented in seminars, conferences, classes, or other academic activities, in which only aggregated results will be mentioned and never individual ones. The researchers responsible for the study ensure that your data will not be processed for purposes other than those previously indicated. The statistical processing database may be eventually required by the scientific journal for the publication of results. In such cases, the statistical processing database will have to be shared with open access, and participants will never be identifiable in any way.

Contacts: To clarify any questions about these studies, you can contact the Director of the Social Psychology Laboratory, Professor Isabel R. Pinto, at the email address (Obermaier, 2022). Specific questions about the treatment of your data can also be addressed to the Data Protection Officer of the University of Porto (dpo@reit.up.pt).

Would you like to participate in this study? By signing/initialing, you indicate that:

- You are 18 years or older; you have read and understood the information above and voluntarily agree to participate in these studies;
- You authorize the collection, processing, and storage of the personal data identified above for the intended purpose and agree with the method of disseminating the results.

Participant's Signature/Initials

Appendix 2

Imagine that, due to a natural catastrophe, life on our planet becomes unbearable and the only hope of survival of the human species is to explore the universe searching for a new place with similar conditions to those that once existed on Earth. You've been traveling for several years in a spaceship with 3000 people, in a distant galaxy, e by chance you discover a planet with high rates of oxygen in the atmosphere, green spaces and water, called Dorothy. Everyone celebrates in contentment because you've found a new home; but unfortunately, as you land, there's a system failure and the spaceship if left severely damaged, killing some members of the crew.

The survivors begin to explore the place where they find themselves, and to their surprise, creatures physically indistinguishable from humans approach them. They quickly realize that on this planet there is a civilization more advanced than the human species - the Dorys - and that they also have cities, transportation, businesses, hobbies, etc. Contact is possible, but limited, since they speak a language similar to Portuguese but with some differences. The Dorys reveal that they normally do not tolerate visits from people from other planets, but due to the similarities between you, they are willing to welcome you. Furthermore, they claim to have a shortage of labor in some industries, so there will certainly be jobs for everyone.

As the days pass, life on the planet Dorothy becomes increasingly difficult. As promised, the Dorys provide jobs for the Earthlings - but with a very heavy workload that barely leaves time for sleep or meals. You are not given days off, even though you begin to realize that the Dorys value leisure and free days in their culture. You are provided with temporary housing since your spaceship is partially destroyed, but the houses are too small, and you have to sleep about 5 people in each room.

In all public spaces, you notice that they look at you suspiciously and strangely at your clothes, customs, and language. One day, a fellow tells you that he was kicked off a bus on his way to work for speaking Portuguese with his wife. He also reveals that they yelled at him that "such languages are not spoken on the planet Dorothy, and if you want to speak like that, you can go back to where you came from." Another earthling tells you that he asked a superior if he could have a day off because he was extremely exhausted and needed rest, to which the boss replied, "You people are ungrateful, we give you everything, and you still come with demands."

Hours later, you see a woman being insulted for being on her knees near the crashed spaceship while praying - the Dorys do not believe in religions.

Situations like the ones described above are becoming more frequent, and frustration among the Earthlings grows. Months later, a small group decides to protest in front of a city hall, demanding decent living conditions. They are met with violence from the police and contempt from the population of the planet Dorothy, who spit, insult, throw stones, and continuously shout, "Go back to your planet," "We don't want you here," "Don't occupy what is ours," among other things. Surprisingly, the mayor is willing to receive you and listen to you, despite the opposition from the general population. After hearing some complaints, the mayor states that you are exaggerating and that the Dorys are naturally hospitable people who have given you everything you need to feel welcome. He advises you not to provoke or irritate them and suggests that you should make a collective effort to fully integrate.

Unfortunately, you cannot repair your spaceship, and the Dorys claim not to have the necessary technology for its repair. All hope of escaping the planet slowly fades away. Tensions between Earthlings and Dorys increase day by day.

Now, reflect on how you would feel on the planet Dorothy.

Appendix 3

Before you start playing, please read the instructions and rules of the University Games carefully:
- You will have 2 minutes to complete each round of questions. After 2 minutes, the game will automatically advance.

The Quiz consists of 3 rounds of questions. Each round will have 10 questions. At the end of each round of questions, you can take a break of about 2 minutes.

For each correct answer, you will be awarded 5 (five) points. If you do not provide any answer, the score will be 0 (zero) points.

If the selected answer is incorrect, the score will be 0 (zero) points, with no penalty (meaning no points will be deducted for incorrect answers).

The use of electronic devices of any kind (computers, tablets, mobile phones, smartwatches, and/or others) to search for the correct answer is not allowed. If any irregularity is detected, the offending player will be disqualified, and the team will be removed from the University Games.

Before the start of the Quiz and at the end of each round of questions (during the break between rounds), team members will have the opportunity, if they wish, to communicate with each other through the team chat. The chat is anonymous and private, restricted to members of each team.

In the team chat platform, there are three buttons: one button is for sending messages to the entire team; another button allows you to contact any other team member directly and privately; another button allows you to anonymously report any irregularity or misconduct by any team member in the team chat. In the event of a misconduct report, the experimenters may access the chat content.

[Continuation of salience of prescriptive norms condition - additional rules]

In the team chat platform, there are three buttons: one button is for sending messages to the entire team; another button allows you to contact any other team member directly and privately; another button allows you to anonymously report any irregularity or misconduct by any team

member in the team chat. In the event of a misconduct report, the experimenters may access the chat content. The University of Porto adheres to an ethical code of academic conduct, and all its members should maintain an "inclusive posture, rejecting and sanctioning any discriminatory, harassing, intimidating, retaliatory, physically violent, or morally coercive practice." Offensive, discriminatory, violence-inciting, or hateful language that jeopardizes the well-being of participants should not be tolerated. Any violation of the code of conduct should be reported by participants.

Appendix 4

Table 6.

Correlation Analysis Between All Bystander Intervention Model Steps Regarding Control x Control Condition

Variable		*M*	*SD*	1	2	3	4	5
1.	Step 1	6.51	0.84					
2.	Step 2	6.56	.73	**.78****				
3.	Step 3	6.56	.78	.63**	**.61****			
4.	Step 4	5.30	1.23	.37**	.27	**.38****		
5.	Step 5 - SR	1.73	.45	.18	.09	.35*	**.40****	
6.	Step 5 - RB	1.41	.50	.26	.29*	.30*	**.47****	.43**

$^{*}p \leq .05$; $^{**}p \leq .01$; $^{***}p \leq .001$

Note: STEP1 = Notice and identify hate content; STEP2 = Interpret the situation as serious; STEP3 = Accept the responsibility to intervene; STEP 4 = Know how to intervene; STEP5 SR = Self-report of the hate speech occurrence; STEP5 RB = Pressing the report button (actual behaviour)

Table 7.

Correlation Analysis Between All Bystander Intervention Model Steps Regarding Perspective-taking x Control Condition

Variable		*M*	*SD*	1	2	3	4	5
1.	Step 1	6.53	.93					
2.	Step 2	6.41	.87	**.70****				
3.	Step 3	6.34	.91	.54**	**.64****			
4.	Step 4	5.41	1.35	.32*	.27*	**.41****		
5.	Step 5 - SR	1.77	.42	.31*	.49**	.36**	.14	
6.	Step 5 - RB	1.53	.50	.48**	.41**	.46**	**.29***	.58**

$^{*}p \leq .05$; $^{**}p \leq .01$; $^{***}p \leq .001$

Note: STEP1 = Notice and identify hate content; STEP2 = Interpret the situation as serious; STEP3 = Accept the responsibility to intervene; STEP 4 = Know how to intervene; STEP5 SR = Self-report of the hate speech occurrence; STEP5 RB = Pressing the report button (actual behaviour)

Table 8.

Correlation Analysis Between All Bystander Intervention Model Steps Regarding Perspective-taking x Prescriptive Norms Condition

Variable		*M*	*SD*	1	2	3	4	5
1.	Step 1	6.55	.81					
2.	Step 2	6.54	.56	**.51****				
3.	Step 3	6.44	.81	.24	**.30***			
4.	Step 4	5.70	1.10	.28*	.14	**.43****		
5.	Step 5 - SR	1.81	.39	-.09	.03	.16	.26	
6.	Step 5 - RB	1.68	.47	-.13	.08	.12	**.41****	.63**

$^{*}p \leq .05$; $^{**}p \leq .01$; $^{***}p \leq .001$

Note: STEP1 = Notice and identify hate content; STEP2 = Interpret the situation as serious; STEP3 = Accept the responsibility to intervene; STEP 4 = Know how to intervene; STEP5 SR = Self-report of the hate speech occurrence; STEP5 RB = Pressing the report button (actual behaviour)

Table 9.

Correlation Analysis Between All Bystander Intervention Model Steps Regarding Control x Prescriptive Norms Condition

Variable	*M*	*SD*	1	2	3	4	5
1. Step 1	6.16	1.24					
2. Step 2	6.00	1.14	**.75****				
3. Step 3	6.23	1.05	.49**	**.51****			
4. Step 4	5.58	1.11	.13	.08	**.32***		
5. Step 5 - SR	1.81	.40	.42**	.38**	.33*	**.32***	
6. Step 5 - RB	1.58	.50	.50**	.56**	.42**	**.48****	.57**

*$p \leq .05$; **$p \leq .01$; ***$p \leq .001$

Note: STEP1 = Notice and identify hate content; STEP2 = Interpret the situation as serious; STEP3 = Accept the responsibility to intervene; STEP 4 = Know how to intervene; STEP5 SR = Self-report of the hate speech occurrence; STEP5 RB = Pressing the report button (actual behaviour)