DOCTORAL (PHD) DISSERTATION

Szekeres Hanna

THE FAILURE TO CONFRONT PREJUDICE

2021

EÖTVÖS LORÁND UNIVERSITY FACULTY OF EDUCATION AND PSYCHOLOGY

Szekeres Hanna

THE FAILURE TO CONFRONT PREJUDICE

Doctoral School of Psychology

Head of School: Prof. Dr. Anikó Zsolnai, Eötvös Loránd University

Programme name: Socialization and Psychology of Social Processes Program

Head of Programme: Dr. Nguyen Luu Lan Anh, Eötvös Loránd University

Supervisors:

Dr. Anna Kende, Eötvös Loránd University

Dr. Tamar Saguy, Interdisciplinary Center (IDC) Herzliya

Prof. Dr. Eran Halperin, Hebrew University Jerusalem

Committee members:

Dr. X, President

Dr. X, Secretary

Dr. Borbála Simonovits (Opponent)

Dr. Zsolt Péter Szabolcs (Opponent)

Dr. X, Member

Dr. X, Member

Dr. X, Member

Budapest, 2021

Table of Contents

List of Figures and Tables

Acknowledgement

I would like to thank for their help and support in preparing this dissertation:

My supervisors, Anna Kende, Tamar Saguy, and Eran Halperin, who were always ready to help me at every step of the way, both academically and personally, and they encouraged me, and contributed immensely to this work.

My insightful friends and colleagues from our past research lab at IDC, where all of this was born.

My friends and colleagues from the Social Psychology department at ELTE for all their spirit and support.

My family for being so supportive and encouraging (and occasionally helping me with Hungarian grammar and style), and to my most insightful and supportive partner, who was very patient with me during the fruitful but not always peaceful writing process.

Abstract

While people generally believe they would stand up for others and act against prejudice and discrimination, historical precedents and empirical evidence suggests that they often fail to do so. In the current research, across seven experiments, I investigated the negative intergroup costs of failure to confront prejudice, focusing on bystanders who are not the target of prejudice, and analyzed potential psychological–moral messages that could motivate bystanders to speak up. We conducted the present studies in the US and in Hungary (N = 1629), in various intergroup contexts where the outgroup minority was either African-, Muslim-, or Latin-American (US), or Jewish (Hungary). For the current research, to test actual confronting, I developed and used across studies an online behavioral paradigm, where participants believed they are witnessing prejudice and discrimination against an outgroup minority and have an opportunity to confront the perpetrator. In this dissertation, I will first review the literature of confronting prejudice (Chapter 1), then I will continue with the first empirical chapter on the motivated prejudice effect (Chapter 2). In this research (N = 922), we tested the impact of failure to confront prejudice on the intergroup attitudes of the bystander. Drawing on cognitive dissonance and self-justification theories, we predicted and found that those who did not confront the perpetrator, albeit having an opportunity to, subsequently endorsed more negative outgroup attitudes compared to their initial attitudes and to control groups – likely in order to justify and reconcile with prior inaction. In this work, we demonstrated a route via which prejudice (not confronted) perpetuates and intensifies in society. In the second empirical chapter (Chapter 3), I present our research on the impact of a moral mindset intervention (N = 707). We tested whether the prospect of moral loss (failure) or moral gain (success) in relation to intervening can motivate people to confront prejudice. Drawing on regulatory focus and loss aversion theories, we predicted and found that a moral loss framing/mindset increases confronting tendencies among those who are morally committed to non-prejudice (possibly due to one's desire to safeguard moral self-concept). Meanwhile a moral gain mindset

had no effect on confronting. In this research, we devised a moral mindset intervention that (a few days later) affected actual confronting, and which can be effectively used in promoting people’s tendency to speak up against prejudice. In the last chapter (Chapter 4), I will discuss the overall findings and their theoretical and applied relevance.

Chapter 1: Main introduction

„I spent so much of my life telling people the things they wanted to hear instead of the things they needed to, told myself I wasn't meant to be anyone's conscience because I still had to figure out being my own, so sometimes I just wouldn't say anything, appeasing ignorance with my silence, unaware that validation doesn't need words to endorse its existence. When Christian was beat up for being gay, I put my hands in my pocket and walked with my head down as if I didn't even notice. I couldn't use my locker for weeks because the bolt on the lock reminded me of the one I had put on my lips ... Silence is the residue of fear. It is feeling your flaws gut-wrench guillotine your tongue. It is the air retreating from your chest because it doesn't feel safe in your lungs. Silence is Rwandan genocide. Silence is Katrina. It is what you hear when there aren't enough body bags left. It is the sound after the noise is already tied. It is charring. It is chains. It is privilege. It is pain. There is no time to pick your battles when your battles have already picked you. I will not let silence wrap itself around my indecision."

(Clint Smith, writer and poet, TED talk, 2014)

Historical events of the 20th century, such as the Holocaust in Europe, genocide in Rwanda, and massacre in Srebrenica, prompted longstanding societal and empirical interest in the phenomenon of people’s failure to stand up for others and intervene in times of racial or ethnic atrocities. While currently in today’s democratic societies such blatant mass tragedies do not transpire, structural and everyday forms of racism still frequently occur and can quickly intensify. For example, consider recent events in this decade when prejudice intensified and lead to the White supremacist “Unite the Right ally” events in 2017 in Charlottesville (in US), the anti-immigrant actions in the aftermath of Brexit in 2016 (in UK), or the serial murders perpetrated by Neo-Nazis against people of Roma ethnicity in 2008-2009 (in Hungary). In our everyday lives, from time to time, we witness someone voicing prejudicial slurs on the bus, at the grocery store, at school, or at the workplace. From time to time, we witness people being discriminated against because of their religion, the color of their skin, or who they choose as their romantic partner. How do you react in these instances? Do you say something? Imagine your co-worker voicing something offensive about another co-worker who is a minority individual, and then discriminating them and not assigning them to an important project. In this instance, you may feel the words coming up from your stomach, in your lung, in your throat – but in the end you decide not to say anything.

In my dissertation research, I investigated the psychological motivations and consequences of witnessing and (failure of) confronting prejudice and discrimination. I focused on bystanders, who do *not* belong to the stigmatized group, but have an opportunity to confront the prejudiced perpetrator, and I investigated (1) the self-justifying harmful consequences of their inaction on their own intergroup attitudes, and related (2) psychological–moral messages that could promote speaking up in face of prejudice and discrimination. In order to test these research questions, we conducted our experiments in two countries, in the United States and in Hungary, across various intergroup contexts, where the outgroup was a racial, ethnic or religious minority. Namely, the outgroup minority was either African American, Muslim American, or Latino (US), or Jewish (Hungary). Given the growth of diverse societies and simultaneous and occasional rise in prejudice (Craig & Richeson, 2014), potential bystanders to prejudice are becoming increasingly common, rendering the focus of the present research timely and relevant. Confronting prejudice is an important socio-political behavior because it provides an opportunity to communicate disagreement with prejudicial treatment within an interpersonal interaction, and to promote an inclusive climate.

In this chapter, I will provide information on the socio-political context of the research (in U.S. and in Hungary), on prevalence of prejudice and discrimination and its effect on the stigmatized groups. I will then review the literature and past research on confronting prejudice. Finally, I will overview and present the research projects of this dissertation.

Socio-political and intergroup context

In my empirical research, I investigated the hypotheses in four intergroup contexts, where the outgroup minority was either Jewish in Hungary (Study 1), or African American (Pilot studies), or Muslim

American (Study 2, Study 4-5), or Latino American (Study 3-4) in the U.S. I did not intend to draw parallels between the status of these minority groups in the different countries, but rather I wished to vary the intergroup contexts across studies to increase external validity of our findings and also to keep in mind what intergroup context was relevant at the time of conducting the studies (more on this below).

In **the U.S**., the cultural and historical background of these stigmatized minorities are quite different, but in today's U.S. they are all considered central in terms of history, politics and societal issues. African Americans comprise around 13%, Hispanic or Latino 19% (U.S. Census Bureau, 2019),[1] and Muslims are 1% of the US population (Pew Research Center, 2017)[2]. Black or African Americans are the largest minority in the U.S., they are largely the descendants of enslaved people who were brought from their African homelands by force in 17th century, and only in 1865 was slavery officially abolished, but they continued being secondary citizens with limited rights up until the 1960s civil rights movement and its following Civil Rights Act, Voting Rights Act, Fair Housing Act. In late 1970s U.S. organizations started a practice called "affirmative action", that is, policies and initiatives aimed at compensating for past discrimination on the basis of race, color, sex, religion or national origin. Among many historical moments to follow, during 2013-2016 the BlackLivesMatter movement started and trended in reaction to police brutality cases, [3] and our pilot studies with African American outgroup was conducted in this period and context. According to research, the predominant negative prejudice about Black people, even today, concerns perceptions about their intellectual and academic ability (e.g., Ashburn-Nardo & Johnson, 2008; Aronson et al., 2002; Devine & Elliot, 1995; Steele & Aronson, 1995), and about criminality especially in regard to Black men (e.g., Oliver, 2003).

[1] https://www.census.gov/quickfacts/fact/table/US/PST045219

[2] https://www.pewresearch.org/fact-tank/2017/08/09/muslims-and-islam-key-findings-in-the-u-s-and-around-the-world/

[3] https://www.britannica.com/topic/African-American/Slavery-in-the-United-States ; https://www.history.com/topics/black-history/black-history-milestones

The American Latinx (new term for Latinos and Hispanics) history is a diverse and long one, with immigrants, refugees and Spanish-speaking or indigenous people living in the U.S. before the nation was even established.[4] This minority group identifies themselves as being of Spanish-speaking background and trace their origin or descent from Central and South America, and other Spanish-speaking countries (or Brazil). The issue of Latinx immigration to the U.S. has been a strong political topic in the second half of the 20th century, and remains so today. Around and following the 2016 U.S. Presidential elections, candidate and later President Donald Trump repeatedly used anti-immigration narratives for political campaigning (e.g., called for a wall between the U.S. and Mexico, for a deportation force to deport all immigrants, and also stated that immigrants from Mexico bring drugs and crime across the border and called them "rapists."). Similarly, this political discourse also targeted Muslim immigrants, for example with Donald Trump (during his presidency) proposing a ban on Muslims entering the U.S. and wanting to suspend immigration from countries with histories of terrorism.[5] Our studies with Latinx and Muslim outgroups was conducted within this political context. In designing our studies, we considered that the predominant negative prejudice about Muslims is based on political bias about terrorism and thus characterized by interpersonal fear and distrust of Muslims (e.g., Kunst et al., 2012; Lee et al., 2013; Oswald, 2005). The predominant prejudice of Latinx people concerns allegedly not paying taxes and abusing the social welfare system (Abad-Merino et al., 2013; Valentino et al., 2013).

In Hungary, Roma and Jewish people can be considered as the main minorities. They suffered the most ethnic hostilities in the 20th century in Hungary, with Roma and Jewish people being victims of the Holocaust. In Hungary, ethnic census is not collected, therefore we only have a rough estimation of the Roma and Jewish population, but it is suggested that Roma people comprise around 7.5% (750,000; European

[4] https://www.history.com/topics/hispanic-history/hispanic-latinx-milestones

[5] https://ballotpedia.org/Donald_Trump_presidential_campaign,_2016/Immigration

Roma Rights Center) and Jewish people comprise around 1% (110,000; Kovacs & Barna, 2018) of the Hungarian population. The Roma are a culturally diverse group with a long history of severe discrimination in all areas of life, marginalization, and poverty (Barany, 2000; Feischmidt et al., 2013; Ladányi, 2001; Pogány, 2006). The Jewish minority in Hungary was mostly annihilated in the Holocaust, and those fewer who survived and did not fled the country afterwards, remained living in Budapest. Following the transition period in 1989, there was a religious and cultural revival of Judaism, and today there is a more active Jewish community, but still primarily within Budapest (Kovács, 2010). There are differences between these groups on the nature of prejudice and discrimination held against them by the majority population, which originates from their different demographic and socio-economic status, cultural identity and history (see e.g., Kovács, 2002; Kemény et al., 2004; Szekeres, 2020). Antigypsyism can be characterized by strong ethnic stereotyping (e.g., regarding work ethic), perception of abusing the social welfare system, and personal aversion (Kende et al., 2017; Ljujic et al., 2012). Meanwhile, personal aversion is less typical of contemporary antisemitism, and prejudice is rather connected to political (ideological) interests (Fábián, 1999; Kovács, 2014). There are popularly held beliefs that Jews have too much control over media, politics, and economics , that is, conspiracy beliefs about secret Jewish world and economic alliances (Bernát et al., 2013; Kende et al., 2018; Kovács, 2010) - altogether reflecting prejudicial beliefs about Jews being manipulative and untrustworthy. For the present research we focused on a Jewish outgroup context primarily because at the time of the study, there was a politically relevant discourse about whether then present governmental campaigns against George Soros tap into anti-Semitic beliefs.[6]

Prevalence of prejudice and its negative impact

Both in the US and in Hungary, discrimination based on race, ethnicity or religion is against the law. In the U.S., the Civil Rights Act in 1964 was a milestone in battling legal discrimination of minorities. Today federal laws prohibit discrimination based on a person's national origin, race, color, religion, disability, sex, and familial status.[7] Laws prohibiting national origin discrimination make it illegal to discriminate because of a person's birthplace, ancestry, culture or language. The Civil Rights Division of the Department of Justice enforces federal laws that prohibit discrimination in education, employment, housing, voting, etc. Similarly, Hungary has ratified most of the major international instruments combating discrimination. The corner stone of the regulation is the general anti-discrimination clause set forth by the Article XV Section (1)-(5) of the Constitution. The general ban on discrimination is further elaborated in the comprehensive antidiscrimination code (’Equal Treatment Law’) that was adopted before Hungary entered into the EU in 2004, and an *Equal Treatment Authority* (hereinafter ETA) was established, which is an autonomous public administrative body with overall responsibility for ensuring compliance with the Equal Treatment Law. Consequently, the ETA deals with discrimination based on age, disability, gender, racial or ethnic origin, religion or belief, sex and sexual orientation, etc. In case of violation of the principle of equal treatment the ETA has broad powers and competences, and it predominantly works as a quasi-judicial body and

[6] Another reason for choosing Jewish outgroup is the difference in social norms about overt expression of prejudice. Openly hostile public discourse is more permitted and typical about the Roma than about Jews (Csepeli et al., 2011; Kende et al., 2018). While having parallel outgroups in Hungary and in the U.S. was not a goal, the norms about prejudice expression regarding Jews in Hungary (especially among university students in Budapest, who participated in the relevant study) seemed closer to the U.S. context. Due to this difference in norms, in my assessment, it was more likely that participants feel (detectable) discomfort for not confronting if Jewish people are insulted than if Roma are – which mechanism was in the focus of the first research.

[7] www.justice.gov/crt/federal-protections-against-national-origin-discrimination-1

often imposes fines. Since the establishment of the ETA, it successfully protected fundamental rights, delivered numerous landmark decisions, and enforced equal treatment by bringing cases to courts and providing assistance, to individuals with various ethnic and religious background.[8]

However, despite the laws, discrimination of minorities occurs in various areas of life, such as in housing, access to health care, employment, education, law-enforcement, or jurisdiction (e.g., Lee et al., 2019 for U.S.; Sik & Simonovits, 2012 for Hungary). Moreover, minorities in both countries frequently experience "everyday prejudice", such as staring, prejudicial slurs, insensitive jokes or microagressions in the form of political discourse (e.g., through media) to interpersonal interactions (e.g., in public areas, workplace). Based on data from the Pew Research Center in the U.S., on average around 75% of Black adults (in 2019) say they have been discriminated against because of their race at least from time to time (incl. 13% who say this happens regularly)[9] ; More than half of U.S. Hispanic adults (58%) say they have experienced discrimination or been treated unfairly because of their race or ethnicity (in 2019)[10] ; And nearly half of American Muslims (48%) say they have experienced at least one of these types of discrimination – being called offensive names, or singled out by airport security, or by other law enforcement officials (in 2017).[11] In Hungary, according to a recent survey conducted among Jews (Kovács & Barna, 2018): (1) Almost 19% stated that they had been verbally insulted or harassed because of their Jewishness personally (in the year prior), and around 27% had witnessed such behavior; (2) 1% of respondents had been victims of physical attack, and 3% had been

[8] Since 1st of January 2021 the Equal Treatment Authority is no longer an autonomous public administrative body, because its tasks and competences were transferred to Commissioner of Fundamental Rights (Hungarian Ombudsperson).

[9] https://www.pewresearch.org/fact-tank/2019/05/02/for-black-americans-experiences-of-racial-discrimination-vary-by-education-level-gender/

[10] https://www.pewresearch.org/fact-tank/2019/07/02/hispanics-with-darker-skin-are-more-likely-to-experience-discrimination-than-those-with-lighter-skin/

[11] https://www.pewforum.org/2017/07/26/findings-from-pew-research-centers-2017-survey-of-us-muslims/#roughly-half-of-muslims-say-they-have-experienced-recent-discrimination

witnessed to such attacks.[12] (3) Almost half (48%) of respondents said they had personally heard verbal antisemitic statements on the street or on public transport, 15% in the workplace, and 10% in government institutions, from authorities, or in their neighbourhood. Finally, (4) more than half of the respondents (55%) said the extent of antisemitism in Hungary is "large", while a further 10% said it is "very large."

Experiencing such prejudice take a toll on the stigmatized individuals in various ways (e.g., Swim et al., 2003; Sue et al., 2007). For one, prejudice and discrimination in employment and in workplace setting affects hiring, and one's professional ambition, advancement, and job satisfaction (e.g., Triana et al., 2015) – thus it has an economic toll on the stigmatized individuals. Additionally, exposure to prejudice has a psychological toll on the person (for review see Barreto & Ellemers, 2015), causing lower self-esteem and self-worth (e.g., Twenge & Crocker, 2002), which not only affect educational and professional performance and achievement (e.g., Nguyen & Ryan, 2008; Walton & Cohen, 2007), but even affects mental and physical well-being and health (for review see Williams et al., 2019; For meta-analyses see Paradies et al., 2015; Pascoe & Smart Richman, 2009; Schmitt et al., 2014). These negative impacts highlight the importance of focusing on methods to negate prejudice expression even in its everyday form. One such strategy is confrontation of people who openly espouse prejudice.

Confronting prejudice

Confronting prejudice means that a person expresses disagreement or disapproval with the prejudicial or discriminatory treatment directly to the source of prejudice (Mallett & Monteith, 2019; Shelton et al., 2006). Building on this definition, in this section, I will first review empirical work on potential motivators for confronting prejudice, the prevalent discrepancy between people's imagined (high) vs. actual (low) confronting rate, potential explanations for this discrepancy, including perceived costs

[12] Note, Jewish people are not identifiable by look as in Hungary they do not wear religious clothing.

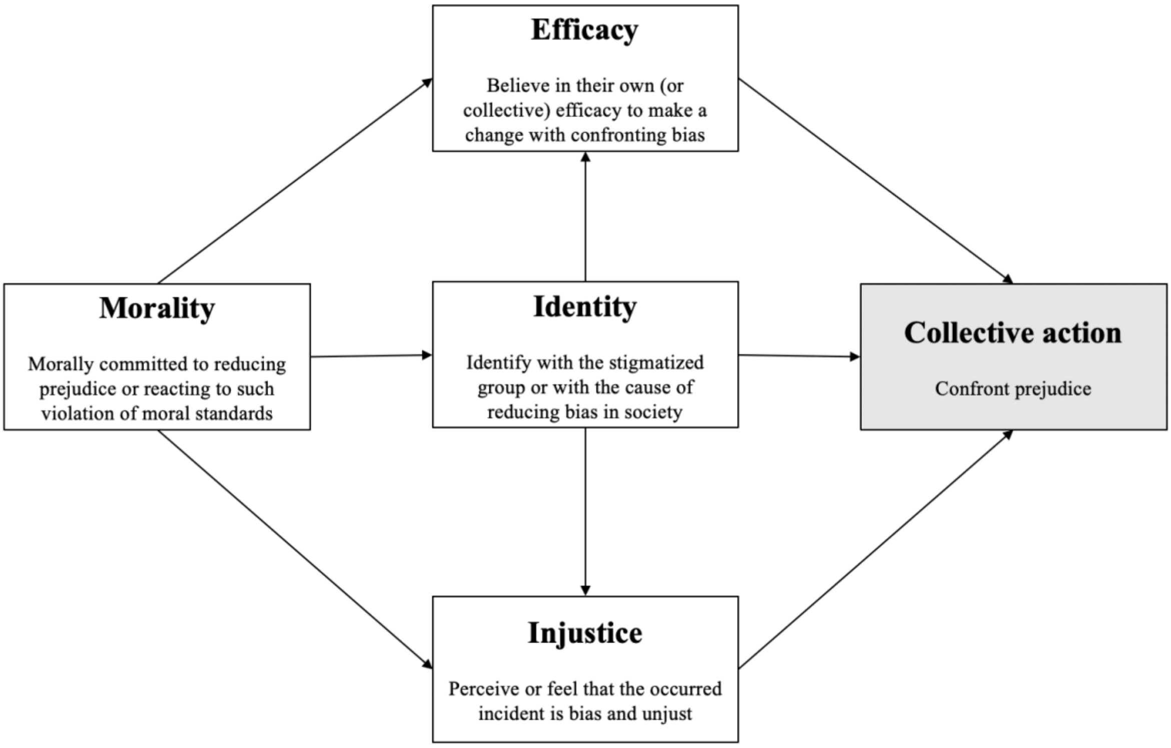
of confronting. Finally, I will present evidence on how people tend to underestimate the cost of inaction, as well as underestimate (or perhaps undervalue) the many benefits of confronting – for example, how confronting can be effective in reducing prejudice in others. During my extensive literature review, where applies, I will point out gaps in the literature, and outline the two research projects that we conducted for the present dissertation.

Ahead I will point out that the majority of the confronting bias literature focuses on targets of prejudice (members of the stigmatized group), and on confronting sexism (sometimes even specifically sexual harassment) and heterosexism. Meanwhile my research focuses on non-target bystanders who confront prejudice based on race, ethnicity and religion. There is likely a difference depending on type of bias confronted (it may involve different norms, personal risks or benefits), and an even bigger difference if the bystander is a target or non-target. Nevertheless, I will include these studies in the review because certain psychological mechanisms are similar, and they provide knowledge to gain a better understanding of the phenomenon in interest.

Motivations for confronting

For outlining people's potential motivations for confronting prejudice, I will use the Social Identity Model of Collective Action (hereafter SIMCA; van Zomeren et al., 2008) as a framework. Collective action is defined as any action that individuals undertake (directly or indirectly) as psychological group members, and with the subjective goal to improve their own group's or another group's conditions (van Zomeren & Iyer, 2009; Wright, 2009). Although confrontation of prejudice had been previously defined as a form of collective action (Munder et al., 2020), previous work has not used SIMCA to thematize motivations for confronting. In empirical work of SIMCA, three main core motivations are identified as predictors of collection action, which is social/politicized identity, perceived injustice (and related emotions), perceived efficacy (van Zomeren et al., 2008). Additionally, later on, moral values were also defined as a core social-psychological motivation to undertake action (van Zomeren, 2013). Therefore, when we consider why individuals confront prejudice, the answer is that they likely identify with the stigmatized group or identify with the cause of reducing bias in society (e.g., Wang & Dovidio, 2017), they perceive the occurred incident unjust (e.g., Ball & Branscombe, 2019), they believe in their own (or collective) efficacy to make a change with confronting bias (e.g., Rattan & Dweck, 2010), and they are morally committed to reducing prejudice or reacting to such violation of moral standards (e.g., Schmader et al., 2012). See Figure 1 for the explained model.

Figure 1. Applying the Social Identity of Collective Action Model (SIMCA; van Zomeren, 2013) to the act of confronting prejudice.


Efficacy
Believe in their own (or collective) efficacy to make a change with confronting bias
Morality
Morally committed to reducing prejudice or reacting to such violation of moral standards
Identity
Identify with the stigmatized group or with the cause of reducing bias in society
Collective action
Confront prejudice
Injustice
Perceive or feel that the occurred incident is bias and unjust

Identity

Prior work finds that identification with one's own group (i.e., how central is the group membership to one's self-concept), or relevant socio-political identification predicts confrontation of bias (Shelton et al., 2006). For example, among women, the stronger they identify with their gender, the more likely they would confront sexism (Munder et al., 2020), even when considering the social costs of confronting (e.g., worry that other people would make fun of them or dislike them if they stand up for themselves; Good et al., 2012). Similarly, among women, feminist identification is also associated with higher likelihood to confront sexism (Ayers et al., 2009; Swim & Hyers, 1999). Further research provided causal link between identification and strengthened confronting intentions (Wang & Dovidio, 2017). In this study, the salience of gender identity was experimentally manipulated and its effect on women's decision to confront a sexist comment was measured in a simulated online interaction with a male partner. Participants who were primed to focus on their gender identity perceived the interaction partner's remarks as more sexist and were more likely to engage in confrontation, compared to female participants who were primed to focus on their personal identity. Note that for non-targets, instead of ingroup identity, it is politicized identity (e.g., identification with a social cause or movement, for example to protect the rights and well-being of minority groups) that can predict undertaking action against the discrimination of other people and groups (van Zomeren et al., 2018), however in the context of prejudice confrontation, this association was not tested in previous research. (But for a similar perspective, see the section of morality below.)

Perceived injustice

Many people confront because they feel the witnessed prejudice is a socially unfair treatment and they disagree with it or it disturbs them (Ashburn-Nardo & Karim, 2019). For example, in a study among high school students, heightened sensitivity to injustice was associated with more engagement in active bystander behavior in response to observing homophobic behavior in the school (Poteat & Vecho, 2016). Additionally, injustice may also motivate confronting through guilt. For example, if white individuals witness another white individual derogating a Black person, they may perceive it as unjust, feel guilty in the name of their group, and to alleviate their guilt, they may confront prejudice (Ball & Branscombe, 2019). Overall, perceiving an act as an “emergency”, that is, gravely unjust, is considered one of the main steps in taking action against prejudicial treatment (e.g., Ashburn-Nardo, 2018).

Perceived efficacy

When people believe confrontation is likely to change perpetrators’ behavior, they are more willing to confront prejudice (Rattan, 2019). For example, women were more likely to report confronting sexism if they believed that the confrontation would yield benefits of making a difference, that is, believing that confrontation would be effective at reducing future instances of sexism, and it would override their concerns about perceived costs of social repercussions (Good et al., 2012). Whether the confronter is generally an optimistic person or believes that it is possible to change the mind of the perpetrator also inserts an effect on confronting willingness. For example, women with a more optimistic outlook on life appraised confronting sexism as more benign. That is, they viewed this process as one that is lower in costs and higher in benefits and were more confident in their abilities to confront sexism (Kaiser & Miller, 2004). Further research showed that optimism increases women’s plans to confront gender discrimination, because expect to have successful outcomes in confronting their perpetrator, such as changing the perpetrator’s mind (Sechrist, 2010). Similarly, people with higher levels of dispositional optimism (vs. lower) are more affected by egalitarian messages and subsequently are more likely to confront a racist act (Wellman et al., 2009). Importantly, Rattan and Dweck (2010) found that targets of bias (ethnic minorities or women) who held an incremental theory of personality (i.e., the belief that people can change) were more likely to confront a person expressing bias (towards minorities or women, respectively), than targets who held an entity theory of personality (i.e., the belief that people have fixed traits). These findings held both when these lay theories were measured or manipulated.

Regarding non-targets, empirical evidence is again sparce, however findings from the literature on allyship indicate that non-stigmatized *advantaged* group allies can acknowledge that they have more control and power in society to change the situation of minorities, and this perceived efficacy can motivate outgroup-oriented collective action (Droogendyk et al., 2016). Relying on this finding, we can assume that perceived efficacy to exert change would motivate confronting prejudice.

Morality

The desire to promote and protect benevolent and egalitarian values can be a strong predictor of confrontation of bias. For example, for both men and women, confronting sexism was predicted by higher communal orientation, that is, how much value they saw in helping others (e.g., “I believe people should go out of their way to be helpful”; Gervais et al., 2010). In prior work, as mentioned before, when non-prejudiced White individuals with optimistic personalities were primed with egalitarian norms (how racism is not OK) it increased their actual tendency to confront an anti-Black racist joke (Wellman et al., 2010). Additionally, in another research it was found that the higher White American participants scored on anti-prejudice views (measured with the internal motivation to respond without prejudice scale, ‘IMS’; Plant & Devine, 1998), the more they self-reported negative affect and exhibited distress-related physiological responses to an observed prejudiced behavior against a Black person (Schmader et al., 2012; for similar results, see Torres et al., 2019).

Discrepancy in hypothetical vs. actual confrontation

Prior studies indicate a complex picture about confronting rates. For one, targets of bias tend to confront bias just as much as tend not to confront it. When capturing reactions to experiencing bias, researchers found that around half of African American college students confronted racist bias (daily diary method;[13] Swim et al., 2003), and also half of college female participants confronted sexism (based on a staged lab setting, Mallett, Ford, & Woodzicka, 2016; based on retrospective accounts, Ayres et al., 2009). On the other hand, among non-targets prior studies indicate that confronting is not necessarily the typical response to expression of bias. In a study relying on retrospective accounts, researchers found that one-third of college students reported to confront racism (Dickter & Newton, 2013). Moreover, using staged lab setting, researchers found that when (straight) people were placed in actual situation of witnessing (anti-gay) bias, not a single participant confronted the perpetrator (Crosby & Wilson, 2015).

Does this indicate that (non-target) individuals do not necessarily see bias or prejudice expression as problematic, and confronting as a viable response? Further research points to no, even non-targets are disturbed by such incidents and generally believe they should be confronted (Crosby, 2015; Kawakami et al., 2019). Firstly, an array of studies demonstrated that advantaged group members report negative attitudes to ingroup members' prejudice or harm directed towards disadvantaged outgroups (Devine et al., 1991; Doosje et al., 1998; Johns et al., 2005; Lickel et al., 2005). At the same time, these attitudes may not necessarily translate into taking action when individuals actually encounter bias or derogation of an outgroup. Previous research demonstrates a disparity between people's anticipated and actual reactions to biased incidents (e.g., Crosby & Wilson, 2015; Kawakami et al., 2009).

For one, even targets of bias generally overestimate their own tendency to intervene when experiencing bias (Good et al., 2019). In one study, female participants were either placed in a situation where a male confederate made sexist remarks, or they solely read about such a scenario (Swim & Hyers, 1999). While the majority (81%) of women in the hypothetical condition believed they would explicitly confront the sexist confederate, only a minority (16%) of those in the actual situation did so

[13] In a daily diary method participants record entries about their everyday lives in a journal, they may be asked to report the experiences in focus as soon after they occur.

directly (and altogether only 45% expressed any direct or indirect displeasure with the comment). In another similar study (Woodzicka & LaFrance, 2001), the majority of female participants (62%) believed they would feel angry and confront a sexist and harassing job interviewer, but when placed in the situation much fewer actually did so (36%). Instead, most of them felt anxious and afraid of retaliation and did not confront the perpetrator directly (Woodzicka & LaFrance, 2001; for similar results with daily diary method see Brinkman et al., 2011). Moreover, when female participants were asked to imagine a scenario where a male job interviewer is making sexually offensive remarks, and participants were specifically reminded of the potential social repercussions of confronting, they still overestimated the likelihood of them challenging the sexist person compared to when women actually experienced the situation (Shelton & Stewart, 2004).

The same pattern seems to apply to non-target bystanders. For example, assessing actual vs. hypothetical confronting of homophobia among straight people, Crosby and Wilson (2015) used a homophobic slur and a hidden camera to record the behavior of participants left alone with the individual who had used the slur. Although about 50% of those who imagined witnessing a homophobic slur reported that they would assertively confront the individual who uttered the slur, no participant who witnessed the slur actually confronted the speaker.

Most notably, Kawakami and colleagues (2009) found that although White Americans anticipated feeling very upset at someone who espouses racial biases, when put in just that situation, they reported little negative emotional reaction and did not take the opportunity to socially reject the racist individual. Although researchers did not measure confronting the perpetrator, per se, they provided insight into the discrepancy between belief and actual reaction to racism. Specifically, participants were randomly assigned to “experiencers” vs. “forecasters” and the former were seated in a room with a Black and a white male confederate. At one point, the Black confederate left the room and gently bumped the white confederate’s knee on his way out. After he left the room, the white confederate either made a racist slur or made no comment. Afterward, participants responded to an emotional distress scale and were asked to choose between the two confederates as partner for an upcoming task. Meanwhile, “forecasters” read about a similar interaction either involving or not involving a racist slur, and they were asked to predict how they would feel if they were in the experiencer’s position and to predict which confederate they would choose as a partner. Results showed that while forecasters reported more emotional distress and were less likely to choose the white confederate as partner in the racist slur condition compared to the control, experiencers showed no significant differences. Within the racist comment condition, forecasters were less likely to choose the white confederate as partner than experiencers (and there was no such difference in the absence of a racist comment). Replicating this study (although again not measuring confronting), they also found physiological and cognitive evidence indicating signs of apathy when being exposed to racism (see Karmali et al., 2017).

Other researchers argued that these findings about reactions to racism are likely moderated by personal beliefs about prejudice (Schmader et al., 2012). In this study, white participants were paired with a Black confederate and together they watched a film depicting two white men having an anti-diversity discussion. The higher participants scored on anti-prejudice views (measured with the internal motivation to respond without prejudice scale, ‘IMS’; Plant & Devine, 1998), the more they self-reported negative affect and exhibited distress-related physiological (cardiovascular) responses to the observed prejudiced behavior (for similar results, see Torres et al., 2019).

Further research (among targets) provides a more nuanced framework to reactions to bias. Specifically, across three studies, female participants’ reaction was investigated when exposed to a confederate, who was allegedly their partner for a task in the study, and who made a sexist remark (Rasinski et al., 2013). They found that women who valued confronting and were given the opportunity to confront, but did not, subsequently made more positive evaluations of the sexist perpetrator than those who had no opportunity to confront. Researchers argued that this occurred as a dissonance-reduction strategy whereby participants were motivated to reduce the inconsistency between their beliefs about confronting offensive behavior and their failure to actually do so. If participants convince themselves that he is “not that bad” then prior failure of not confronting would be more in line with the values they place on this behavior (Rasinski et al., 2013). Furthermore, they also found that participants who initially valued confronting but did not confront a sexist perpetrator reduced the amount of importance they placed on confronting socially inappropriate behavior (in general). In further supporting their dissonance-reduction argument, Rasinski and colleagues (2013) also found that when participants were given a chance to affirm an important aspect of the self (meaning re-affirm their integrity after not acting), the subsequent inflated evaluations of the confederate did not occur. A recent study replicated these effects and found that women who did not confront sexism trivialized sexual harassment compared to those who confronted (Mallett et al., 2019; Study 1).[14]

Indeed, the inconsistency between beliefs about how one should react to bias and one’s actual (non-confronting) behavior can give rise to psychological discomfort. Women who generally believed they should confront sexism and were made to think of instances in which they failed to do so, experienced guilt, regret and obsessive rumination (Shelton et al., 2006). Similarly, women reported more dissonance (e.g., “I feel a little conflicted about how I responded”) when they imagined ignoring a sexist remark compared to having no chance to confront it (Mallett et al., 2019; Study 2). Those who are not the target of bias can experience similar feelings. For example, White Americans who felt they should not behave in a prejudiced manner towards minorities but were made to consider how they might actually do so, also experienced discomfort (Voils et al., 2002; Zuwerink et al., 1996).

[14] In this study there was no control group or initial attitude measures, therefore it is difficult to conclude causality, such as whether confronters were to begin with less tolerant of sexual harassment, or they became less tolerant following confronting.

Cognitive dissonance theory asserts that when people behave in ways that are contradictory to their norms, values, attitudes or beliefs, they tend to experience discomfort (Cooper & Fazio, 1984; Festinger, 1957; Stone & Cooper, 2001). This psychological discomfort motivates individuals to employ strategies for reducing the dissonance by changing one of the elements causing the dissonance (e.g., changing the relevant belief that contradicts the behavior) or adding a cognition that helps reduce the overall level of inconsistency (Festinger, 1957). It is proposed that *inaction* can also lead to cognitive dissonance effects (Aronson & Carlsmith, 1963; Tykocinski et al., 1995). For example, in a study, when participants failed to act cooperatively in a social dilemma, subsequently came to justify their inaction by decreasing their perceptions of the likely effectiveness of having cooperated (Kerr & Kaufman-Gilliland, 1997). Be it action or inaction, if the initial counter-attitudinal behavior cannot be changed, people will likely alter their attitudes and views instead (Abelson et al., 1968; Duval & Wicklund, 1972; Festinger, 1957).

In line with dissonance theory, it is possible that the physiological (cardiovascular) discomfort detected by Schmader and colleagues (2012) was an indicator of a dissonance arousal originating from lack of speaking up in face of prejudice. Such dissonance arousal could then motivate a self-justifying dissonance-reduction process that (if successful) would eventually result in emotional indifference – such as the indifference found by Kawakami and colleagues (2009) and Karmali and colleagues (2012).[15] Similarly, their findings regarding the lack of rejection of the racist perpetrator may be also a product of dissonance-reduction whereby participants attempted to self-justify and minimize the severity of the situation in order to reconcile with their lack of reaction to the response and be able to get on with their day. This interpretation of results would align with the findings of Rasinski and colleagues (2013) on the positive

[15] Indeed, Karmali and colleagues (2012) measured physiological signs ~20 minutes after the incident, which can mean that by that time they went through a dissonance reduction process.

evaluation of the sexist perpetrator and devaluation of the importance of confronting.

Drawing on prior work reviewed above, in our first research described in Chapter 2, we apply the logic of cognitive dissonance reduction to non-target bystanders' inaction in face of prejudice. Specifically, we propose that when bystanders witness prejudice and have an opportunity to confront, but do not, they will be motivated to change their attitudes both about the stigmatized outgroup and about the witnessed incident (specifically, *trivialize* it)[16] in order to obtain consistency between their beliefs and their inaction. To reconcile with prior failure to confront, they would rationalize it and convince themselves that the prejudice they observed was based on a reasonable judgment, which would lead them to endorse more negative outgroup attitudes. Whereas prior work focused on the targets of bias (specifically women; Rasinski et al., 2013; Mallett et al., 2019), in the present research we propose that such dissonance can also occur among observers not belonging to the target group. This shift in focus enabled us to go beyond evaluations of the perpetrator (and appraised value of confronting), to assess the devastating cycle of rising prejudice as a function of failure to confront prejudice.

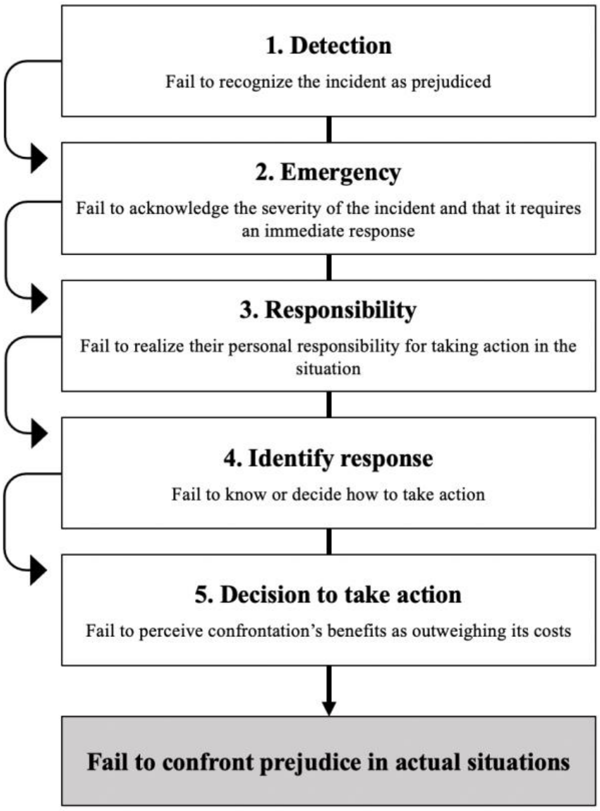
Explanations for discrepancy in confronting intentions

Given that people generally believe they would and have the intention to confront prejudice, the question is warranted: Why do they eventually fail to do so? In order to review the literature that explains this discrepancy and behavioral inaction, I will use the *Confronting Prejudiced Responses* model (hereafter CPR; Ashburn-Nardo et al., 2008). CPR outlines the factors that predict the likelihood that people will confront prejudice, but even more so it highlights the various obstacles that could stand in the way of confrontation even for well-intentioned and motivated individuals (Ashburn-Nardo & Karim, 2019). The CPR is based closely on

[16] In the dissonance literature, trivialization refer to minimizing the significance of the element causing dissonance, and besides attitude change it is a frequently employed dissonance-reduction strategy (e.g., Simon et al., 1995). See further details in chapter 2.

the classical model of bystander intervention in physical emergencies (Latané & Darley, 1970). The CPR proposes that in order to confront prejudice, bystanders: (1) need to recognize the behavior as prejudiced, (2) perceive it as an emergency that requires an immediate response, (3) feel personal responsibility for taking action, (4) identify a response, and finally (5) take action (see benefits of confronting as outweighing the costs). See Figure 2 for the explained model.

Figure 2. Applying the *Confronting Prejudiced Responses* model (CPR; Ashburn-Nardo et al., 2008) to explain the discrepancy between imagined vs. actual confronting rate.


1. Detection
Fail to recognize the incident as prejudiced
2. Emergency
Fail to acknowledge the severity of the incident and that it requires an immediate response
3. Responsibility
Fail to realize their personal responsibility for taking action in the situation
4. Identify response
Fail to know or decide how to take action
5. Decision to take action
Fail to perceive confrontation's benefits as outweighing its costs
Fail to confront prejudice in actual situations

(1) Detection

In order to confront prejudice, one needs to identify a statement or behavior as prejudiced (Ashburn-Nardo et al., 2008). For example, women were significantly more willing to confront overt (vs. subtle) forms of sexist discrimination (Lindsey et al., 2015). That is, when a situation is clearly prejudiced, and there is no ambiguity around it, people are more likely to recognize it as prejudice, which heightens the likelihood of reacting to it. Bystanders may also deter from confronting prejudice that is ambiguous to avoid mistaking a situation, confronting wrongfully, and then loosing face. This fear is justified given that confronters are evaluated negatively when there is ambiguity over whether the confrontation was warranted. For example, Zou and Dickter (2013) found that white participants evaluated a Black target more negatively for confronting a more ambiguous compared to a less ambiguous racist comment (and this perception was particularly pronounced among participants high in colorblind ideology (i.e., the belief that one's race should be ignored, for better or worse). If a comment is ambiguous and perceived as non-prejudiced, confrontation can be seen as overreaction.

However, for members of non-stigmatized groups, who often lack experience of discrimination, recognizing more (or even less) subtle forms of prejudice may be challenging (Ashburn-Nardo et al., 2008). Additionally, non-stigmatized high-status group members may have various motivations for not seeing discrimination (Crosby, 2015), such as to avoid a threat to their ingroup's image (Sommers & Norton, 2006), or due to endorsement of a colorblind ideology (e.g., when whites believe they should not "see" race, they are less likely to recognize racism and to react effectively to instances of discrimination; Apfelbaum et al., 2012; Zou & Dickter, 2013).

Regarding the discrepancy between actual vs. hypothetical confronting, we should firstly consider that these ideologies and group image protection might not be activated for "forecasters" therefore they are not able to account with them when imagining a hypothetical situation. Furthermore, when "forecasters" are asked to predict their behavior, the fact that the question is raised likely already signals that they are questioned about a clearly inappropriate (prejudiced) incident – so detection is easier. When placed in the actual situation, it is more difficult for "experiencers" to determine what is happening exactly, as outlined before.

(2) Emergency

Even if people identify a treatment prejudiced, they may not consider it harmful enough to warrant intervention. To motivate confrontation, one needs to assess the incident as so severe that renders immediate reaction, like an emergency (Ashburn-Nardo et al., 2008). For example, the same sexist or heterosexist sentiments are generally taken less seriously and perceived as less "confrontation-worthy" when delivered as jokes versus as serious statements (Katz et al., 2021; Mallett et al., 2016). Sometimes non-stigmatized individuals seek information from the stigmatized group to decide if a treatment was harmful. For example, in a notable study, when white participants were together with a white and a Black confederate, and they observed the white confederate make a remark that was ambiguously racist, the participants literally turned their eyes to the Black person, as if to determine whether harm was done (Crosby et al., 2008). This is defined as social referencing, that is, seeking out the responses of a potentially victimized group member to help assess the situation (Crosby, 2006). Moreover, a reason why non-stigmatized individuals may discount severity of prejudiced incidents is because they lack information about the personal and damaging consequences of prejudice on the target individuals (Crosby, 2015). Similarly, to detection, when considering the discrepancy of confronting intention, it may be easier for forecasters (than experiencers) to determine whether a situation is harmful, simply because it is being asked from them.

Another important determinant of whether a treatment will be interpreted harmful enough to warrant confronting relates to the bystander's own prejudice toward the targeted group. Since we are considering explanations for discrepancy between hypothetical vs. actual confronting, the question is not really about explicit bias, but about implicit bias. Aversive racism theory suggests that while people are consciously motivated to be egalitarian, they harbor unconscious prejudicial attitudes which direct their judgment and behavior (Gaertner & Dovidio, 1986). In this vein, it is suggested that while "forecasters" predict that they will react in accordance with their consciously held attitudes, when placed in the actual situation, “experiencers” react based on their unconscious biases and thus might not be as offended by the prejudiced treatment and instead they might trivialize it (Kawakami & Karim, 2019).

Furthermore, affective forecasting theory suggests that one potential explanation for inaccurate predictions is the tendency to overestimate the impact of the focal event on our thoughts and feelings, while underestimating the impact of other events occurring during the forecasted incident (Wilson & Gilbert, 2005). When people are asked to forecast their emotions, they tend not to fully realize the impact of other ongoing and upcoming events in their lives (e.g., I offended a friend yesterday) and therefore overestimate the impact of the forecasted incident on their feelings (Kawakami & Karim, 2019). Furthermore, people generally fail to predict the power of social influence on determining their behavior and action. When you need to predict your behavior in a prejudiced situation you would not calculate with other bystanders’ behavior (but instead believe to be irrelevant to your behavior). However when you are placed in the actual situation, on the one hand, you use others’ behavior as cues for the desired action, and also you may simply conform to others’ behavior (Cialdini & Goldstein, 2004) – and if others are not confronting, you likely will neither.

(3) Responsibility

The more bystanders feel personal responsibility for addressing prejudice, the more likely they are to confront it (Ashburn-Nardo et al., 2008; Ashburn-Nardo et al., 2014). For example, people in responsible roles, such as leaders in organizations feel more responsible for confronting prejudice compared to those who are not in such authority roles (Ashburn-Nardo et al., 2020). Additionally, in respect to diffusion of responsibility, women were more likely to confront a man who made a sexist remark when they were the only woman present than when other women were also present (although note that those other women were confederates who did not react; Swim & Hyers, 1999). Similarly, among men, recent research found that pluralistic ignorance about sexism (misperception of group attitudes) can inhibit confronting sexism (De Souza & Schmader, 2021).

Social norms are also likely to determine whether people, especially those not targeted, will feel that speaking out is even a social responsibility or not (De Souza & Schmader, 2021), and many people may feel that only members of affected groups are entitled to respond to discrimination (Crosby, 2015). In regard to discrepancy and norms, forecasters may consider injunctive norms (what people should do) and make decisions based on which reactions are considered socially appropriate (like you should confront prejudice). In contrast, experiencers may attend more to descriptive norms (what most people do) and respond according to what other people would do in the situation (Kawakami & Karim, 2019). If so, this phenomenon becomes a snake biting its own tail: most people will decide not to confront despite accepting the norm that confronting is desirable, because they observe that others do not confront (who based their decision on the same observation). This provides yet another reason why confronting is important – to redefine descriptive norms.

(4) Identifying response

Even if people consider an incident harmfully prejudicial and even feel a sense of responsibility to act, they may still not confront – because they do not know how to (Ashburn-Nardo et al., 2008). This is why when bystanders learn confrontation patterns and practices, it helps them confront. In a study, students were more likely to confront bias experienced in school when they previously attended a workshop where they practiced confronting prejudiced remarks (Plous, 2000). Indeed, another possible explanation for the discrepancy of confronting is that in the heat of the moment it is difficult to decide upon an appropriate response (Ashburn-Nardo et al., 2008). Similarly, the sense of effectiveness of one’s response is also important. To the extent that people believe confrontation is unlikely to change perpetrators’ behavior, they are less likely to report confronting prejudice (Good et al., 2012; Rattan, 2019). For example, in a study mentioned before, Rattan and Dweck (2010) found that targets of prejudice (ethnic minorities or women) who held an incremental theory of personality were more likely to confront a person expressing prejudice (towards minorities or women, respectively), than targets who held an entity theory of personality.

(5) Decision to take action

Finally, confrontation depends on the decision to confront. On the one hand, this decision is determined by all previous factors and steps that were mentioned, namely, motivations, detection, appraised emergency, appraised responsibility, and ability to identify a response. In addition, the CPR model asserts that this last step of decision making essentially relies on the bystander's cost-benefit analysis, whether they perceive confrontation's benefits as outweighing its costs (Ashburn-Nardo et al., 2008). The next and last sections of the literature review are dedicated to the vast literature examining people's perceived and real costs, and benefits, to confronting prejudice.

Costs of confronting prejudice

When (stigmatized or non-stigmatized) bystanders decide to confront, they anticipate potential costs that this action can entail. Given a certain situation, the person confronting prejudice may anticipate or actually risk interpersonal and social (e.g., rebukes, antagonism, ostracism), economic (e.g., job dismissal), physical (e.g., getting hurt) or psychological–mental and emotional (e.g., stress, cognitive load) costs for confronting. I will overview these potential perceived and actual costs below.

Interpersonal and social costs

The most documented barrier to confrontation is anticipated interpersonal and social costs. Stigmatized individuals may avoid confronting prejudice targeted towards them or their group due to self-presentational concerns and fear of social repercussions and retaliation

(Shelton & Stewart, 2004; Swim & Hyers, 1999; for review see Barreto & Ellemers, 2015). For example, a study showed that stigmatized groups (African Americans and women) were less willing to attribute negative events that occur to them to discrimination when they are in the presence of members of a nonstigmatized (vs. their own) group (Stangor, Swim, Van Allen, & Sechrist, 2002). Such tendency occurs because stigmatized individuals are afraid to be labeled as “crying prejudice” or “playing the race or sex card” (Kaiser & Major, 2006). More specifically to confronting, women reported that concerns about social sanctions or disparagement play a significant role in their decision to confront sexism, such as fear that the sexist person or other people would make fun of them or dislike them, or that the sexist person would get upset or angry (Good et al., 2012). Note that potential confronters may fear backlash not only originating from the perpetrator, but also from others who witness or know about their confrontation, or by the broader society.

Such concerns are not unreasonable as people indeed often judge target group confronters unfavorably (Zou & Dickter, 2013). For example, women who blame negative outcome (e.g., receiving a bad grade) on sexism are evaluated as a complainer and disliked, even by members of their own ingroup (Garcia, Reser, Amo, Redersdorff, & Branscombe, 2005). Similarly, African Americans who blame negative educational or professional outcomes (e.g., not getting hired for a job) on discrimination are evaluated negatively and seen as “complainers” and “hypersensitive” (also as irritating and trouble making), even when the discrimination is blatant (Kaiser & Miller, 2001; 2003). In another study, participants read about a woman who confronted a man making a sexist remark, and male (not female) participants’ disliked the target woman less when she confronted the sexist remark than when she did not confront it (Dodd et al., 2001). If stigmatized individuals are aware of the potential negative reception in response to their confronting, such anticipated costs can deter them from taking action against bias.

Some of these costs of confronting might be attenuated for those who are not the targets of bias (Czopp, 2019). When a non-target individual confronts bias, compared to a target, they are evaluated less negatively, their views are taken more seriously, and their actions are seen as legitimate efforts to combat prejudice (for reviews see, Czopp, 2019; Drury & Kaiser, 2014). For example, Czopp and Monteith (2003) investigated how people would react when they were made aware of their gender- or racial-biased responses by another person. They found that when confrontation came from a non-target confederate, compared to a target (White vs. Black, woman vs. man), participants felt more guilt for their bias and meanwhile they were less irritated by the confrontation (although some of these effects were not replicated in Czopp et al., 2006). Similarly, participants who were confronted about their own implicit racial bias perceived a white (compared to the Black) confronter less as a complainer, and in turn, they were also more likely to accept the nontarget confrontation as convincing and suggestive of their personal need to work on bias reduction (Gulker et al., 2013).

Rasinski and Czopp (2010) also investigated how third-party perceivers (and not the perpetrator) evaluate confronters. In their study, white participants watched a scenario where a white person expressed ambiguously racist comments and was either confronted or not by a white or Black person. The nontarget's confrontation was rated as more persuasive in regard to how biased the statement was, while target's confrontation was rated as rude and increased participants' agreement with the initial biased response they listened to (for similar results see Schultz & Maddox, 2013). In the context of confronting sexism, these mechanisms are somewhat more complex (Czopp & Monteith, 2003). For example, Gervais and Hillard (2014) found that third-party perceivers indeed evaluated male confronters of sexism more favorably than female confronters, but only when they confronted subtly in a public context, compared to when explicitly and in private. In another study, also among third-party perceivers, Eliezer and Major (2012) found that while both male and female bystanders who claimed discrimination on behalf of a female coworker were evaluated more negatively than those who did not claim discrimination in the same situation, female bystanders who claimed discrimination were derogated (as complainers) more than male bystanders who did the same.

In the latter study, the found mediating explanation for this effect was that target’s confrontation threatened participants’ beliefs about the fairness of group (gender) status differences to a greater extent than non-target’s confrontation (Eliezer & Major, 2012). The explanation for why non-targets incur less interpersonal costs and backlash is possibly because of perceived lower self-interest and higher altruism (e.g., Czopp et al., 2006; Drury & Kaiser, 2014). That is, people who witness confrontation likely consider that non-targets must be reacting because they objectively perceive the situation as unfair and confrontation-worthy, and target confronters are seen as more subjective and perceived to be acting out of self-interest or group-interest (i.e., they are perceived as overly sensitive and only complaining; Czopp et al., 2006; Eliezer & Major, 2012; Kaiser et al., 2009; Rasinski & Czopp, 2010). Perhaps sensing this attenuated reactions, non-targets consider social costs less so than targets when deciding to confront. For example, men’s, compared to women’s, appraisal of the potential cost of confronting was not a significant predictor of the frequency with which they confronted sexism (Good et al., 2018).

Non-targets may incur less costs for confronting than targets, but they can still receive more negative judgment for confronting compared to not confronting at all (Czopp et al., 2006; Kutlaca et al., 2020). Additionally, non-target confronters may also incur social costs that are less relevant to targets, namely exclusion from their in-group and association with the stigmatized group. Supporting stigmatized outgroup individuals may suggest to others that they are traitors or that they have taken on the characteristics of the stigmatized group. For example, men who challenge sexist actions may be seen as less of a man. Studies also show that heterosexual male allies fear of stigma by association, namely that they will be perceived as gay when confronting anti-gay prejudice (Kroeper et al., 2014), and indeed, heterosexual confronters are perceived by others as possibly gay (Cadieux & Chasteen, 2015).

Economic cost and perpetrator's power status

Another significant factor that people may consider when deciding to confront prejudice is how much power the perpetrator has over them, and whether they would be penalized for speaking up (Glasford & Pratto, 2014). One particular setting where this is a frequent concern is in the workplace context (Ashburn-Nardo et al., 2008), where one could typically fear economic and professional costs for confronting. For example, women interviewing for a job were less likely to confront a sexist interviewer when confronting implied higher costs (i.e., being interviewed for a highly desirable job) than when confronting implied lower costs (i.e., the interview was only to gain experience; Shelton & Stewart, 2004). Women or minorities are less likely to confront sexism or racism, respectively, if the person expressing bias is someone who has power over them (e.g., if that person is a superior, like a boss) and they anticipate negative consequences, compared to when the person has an equal status (e.g., a friend or a co-worker; Ayres et al., 2009; Ashburn-Nardo et al., 2014).

Physical threat

In most contexts where confronting of prejudice was empirically investigated physical threat was less of a risk. Yet, even if there is not much evidence to it, it is safe to assume that for example, if a person is sitting on a bus and overhears a skinhead voicing racist slurs, the fear of physical retaliation for confronting is realistic. For example, in a different context, a study showed that women were less likely to confront unwanted sexual attention than sexist comments or unfair treatment, presumably because of greater threat to their physical safety (Ayres et al., 2009).

Intrapersonal psychological costs

Finally, although less discussed than other threats, the prospect of cognitive and emotional taxing deters people from speaking up against prejudice. For example, Hyers (2007) found that people often avoid confronting because of anticipated expenditure of cognitive and emotional resources (e.g., "The perpetrator wasn't worth my time and energy" or "It would have been emotionally too draining"). Some, especially stigmatized individuals may perceive that confrontation is simply not worth the effort and the emotional investment, especially if they feel they cannot change the mind of the prejudiced person (Rattan, 2019).

Underestimating the benefits of confronting

People have an array of legitimate reasons not to confront, such as the costs that were outlined above, but people may realize that there are also costs of *not* confronting. Firstly, it entails intergroup and societal costs, because ignoring prejudicial treatment increases the likelihood for the behavior to continue unchanged (Czopp, 2013), and inaction communicates that the witnessed prejudicial treatment is appropriate and condoned (Czopp, 2019), both to the broader audience (Blanchard et al., 1991) and also to stigmatized individuals (Hildebrand et al., 2020). Secondly, there are interpersonal costs, such as the loss of respect of others (Mallett & Melchiori, 2019). Lastly, there are intrapersonal psychological costs, such as feelings of discomfort (Mallett et al., 2019; Rasinski et al., 2013; Shelton et al., 2006). Confronting is not only an antidote against these costs, but it can also be beneficial. Those who endorse anti-prejudiced values or perceive a situation as particularly unfair may be willing to forego the negative interpersonal costs (e.g., being disliked by the confronter) knowing that they may reap intrapersonal, interpersonal and social gains (e.g., Good et al., 2012). These benefits of confronting will be reviewed below.

Societal and intergroup benefits of confronting

The primary potential of confronting prejudice for intergroup and societal benefit is to challenge the views of the perpetrator, and of those who are directly or indirectly part of the situation. Overall, according to prior research confrontation can be an effective strategy in reducing prejudice in others (for a review, see Mallet & Monteith, 2019), and it is especially persuasive if done by non-target individuals (Czopp &

Monteith, 2003; Drury & Kaiser, 2014; Gulker et al., 2013; Rasinski & Czopp, 2010; Schultz & Maddox, 2013).

Specifically, regarding the impact on perpetrators, once reproached for their racist or sexist behavior, they tend to feel guilty and become wary of offending again (Monteith et al., 2019). Most notably, in a study by Czopp and colleagues (2006), white participants who were confronted by a confederate for making a racist inference about a Black person, although evaluated unfavorably the confronter, they also felt negative self-directed affect, for example being angry at themselves and feeling guilty. This in turn reduced the likelihood of prejudicial responses in a subsequent (experimental) task, and this was true regardless of the tone (less or more hostile) or the source (Black or white confronter) of confrontation (Czopp et al., 2006).[17] Similarly, when participants made prejudicial inferences about pictures of Native Americans, confrontation by a confederate (compared to no confrontation) reduced participants' future biased responses (Lewis & Yoshimura, 2017). Importantly, this effect of confronting on reduction of prejudice is found to be long-lasting (Burns & Monteith, 2019; Chaney & Sanchez, 2018). The general effect is also found in the context of confronting sexism (e.g., Burns & Granz, 2020), although in some instances it is (again) more complex. In one study, male participants who were confronted about their gender bias by a female confederate (in a face-to-face interaction) later engaged in apologetic, compensatory behavior that increased mutual liking and, in turn, reduced men's use of sexist language in a subsequent conversation (Mallett & Wagner, 2011). In another study researchers showed that providing concrete evidence with claims of bias enhances the impact of confrontation, specifically, participants were confronted with evidence that they evaluated a female applicant for a lab manager position more negatively than an identical male applicant, which confrontation activated greater guilt and, in turn, concern about expressing and regulating gender

[17] Czopp and Monteith (2003) found similar results, but only for confrontation about racial bias (and not sexism), and only among low-prejudiced participants.

bias in the future, compared to when no concrete evidence was provided (Parker et al., 2018).

Confrontation is not only effective in educating the perpetrator, but also in impacting the surrounding social environment, the third-party observers who are directly or indirectly present (Czopp, 2019). In general, simply being exposed to prejudice results in worse perceptions about an outgroup through desensitization or through persuasion and increase in prejudiced norm (Blanchard et al., 1991; 1994; Greenberg & Pyszczynski, 1985; Fasoli, et al., 2016; Krolikowski et al., 2016; Simon & Greenberg, 1996; Soral et al., 2017). However, confrontation of bias can stop this negative impact and even cue positive attitude change. Indeed, the mere act of seeing a biased behavior or statement confronted can reaffirm observers' anti-prejudiced values (Czopp, 2019). For example, previous research studied the effects of normative influence on reactions to racism and found that when white college student participants observed their peer (even one single individual, regardless of whether they were white or Black) expressing strong disapproval of racism (e.g., confronting a racist person) it reaffirmed participants' egalitarian values, and they subsequently expressed more anti-prejudiced opinions (Blanchard et al., 1994). These effects are even more pronounced when there are unclear social norms about the acceptability of prejudice toward a group (Zitek & Hebl, 2007). Thus, confronting can set a social norm about how expressing prejudice is not appropriate.

Beyond communicating an anti-prejudiced norm, confrontations may help signal and define what treatment should be recognized as prejudicial, particularly when it is ambiguous (e.g., a joke) – so that others can respond accordingly. This way confrontation can increase witnesses' evaluation of the initial offending behavior as biased. For example, in a study mentioned earlier (Rasinski & Czopp, 2010), white participants watched a scenario where a person expressed ambiguously racist comments, and when he was confronted by a white confederate, participants perceived the perpetrator as more prejudiced and were less likely to agree with his biased comments compared to when no confrontation occurred. Similarly, college student participants evaluated a hypothetical male student who made sexist comments about women as more offensive and prejudiced when he was confronted (by either a teacher or another male student) compared to when no confrontation occurred (Boysen, 2013). In a similar vein, individuals who expressed heterosexist bias were respected less by participants when they were confronted compared to when they were not confronted (Dickter et al., 2012).

Finally, when prejudice is confronted, it can have a positive impact on targeted individuals (Hildebrand et al., 2020). For example, in a study, when men suggested that sexism had taken place, female participants reported more self-confidence (less self-handicapping and higher personal performance state self-esteem), and they were more likely to file a complaint about it, than when sexism was suggested by a female source (Cihangir et al., 2014). These are important findings, because even if we cannot change the perpetrator's mind it may be comforting to know that with confronting, we can reaffirm an anti-prejudiced social norm or standard for bias.[18] The confronter's public declaration of the unacceptability of bias may serve a broader goal of establishing norms of fairness that goes beyond the confronter-confrontee dyad (Monteith et al., 1996).

Interpersonal and intrapersonal benefits of confronting

When deciding to confront, people can sometimes make a cost-benefit analysis about wanting to be liked or respected (Mallett & Melchiori, 2019). Namely, in some instances, people may decide to sacrifice potential backlash to challenge prejudice if knowing they will receive social support elsewhere, maybe even earn (some) others' respect and admiration (Mallett & Melchiori, 2019). For example, research showed that women who preferred to be respected rather than liked, were more inclined to confront sexism in a staged job interview (Mallett & Melchiori, 2014).

[18] Similarly, for positive effects of others' collective action on one's own attitudes, see Szekeres, Shuman, & Saguy, 2020.

This benefit analysis resonates with reality, as in given situations confronters can indeed gain the respect of others. For example, women liked and respected a woman who directly confronted a male perpetrator's blatantly sexist comments more than a woman who ignored the comments (Dodd et al., 2001). Similarly, Black and Asian participants who strongly (vs. weakly) identified with their racial group evaluated ingroup members who confronted racism more favorably than ingroup members who did nothing (Kaiser et al., 2009).[19] Regarding non-target confronters, when reading about confrontations of racism or heterosexism, non-target participants evaluated majority group members (whites, heterosexuals) who confronted (assertively or unassertively) as more likable, respectable, and moral than those who did not confront at all (Dickter et al., 2011).

Not completely unrelated, but beyond potential interpersonal benefits, confronting can also entail intrapersonal benefits, for example, a sense of empowerment (Czopp, 2019). For example, in a study, college students were exposed to a sexist statement during a staged, online interaction, and confronting was associated positively with competence, self-esteem, and empowerment for women, although not for men (Gervais et al., 2010; see also Hyers, 2007). In this respect, confronting may serve as an antidote for some of the adverse psychological outcomes that stigmatized individuals experience as targets of bias, for example, it may restore a sense of control (Swim & Thomas, 2006). Some studies go as far as to suggest that confronting discrimination is an active coping strategy that buffers against negative health outcomes of experiences of prejudice (Chaney et al., 2015). Regarding non-targets, there is not much empirical evidence about intrapersonal costs/benefits of confronting prejudice specifically. In one study using diary retrospective accounts, heterosexuals who confronted anti-gay bias reported to feel more satisfied with their responses than those who did not confront (Dickter, 2012). Based on this finding, we can assume that there are also benefits for non-targets, in line

[19] They did not find the same pattern for high-identifying women and sexism, but that might be due to the complex relationship between gender identification and feminist attitudes among women (Saguy & Szekeres, 2018).

with benefits identified within the broader notion of intergroup helping or allyship, such as empowerment, the restoration or reinforcement of positive moral self and ingroup image (Droogendyke et al., 2016; Radke et al., 2020).

Intrapersonal cost becoming intergroup cost (or benefit)?

According to prior research, as I mentioned it earlier, when women fail to confront sexism, it can generate intrapersonal psychological costs, such as feelings of guilt and obsessive rumination (Mallett et al., 2019; Shelton et al., 2006). Moreover, women who initially valued confronting and were given the opportunity to confront, but did not, subsequently made more favorable evaluations of the sexist perpetrator and also devalued confronting behavior in general, consistently with self-justifying theories (Rasinski et al., 2013).

Evidence regarding what non-targets experience when they fail to confront prejudice is much more limited. Most studies investigated reactions to *merely* being exposed to a prejudicial situation, in which some researchers suggest that non-targets experience no psychological discomfort whatsoever (while also do not engage in interpersonal rejection of the perpetrator; Karmali et al., 2017; Kawakami et al., 2009) while others suggest that at least low-prejudiced individuals experience psychological discomfort when witnessing prejudice (Schmader et al., 2012; Torres et al., 2019). Prior work has not investigated the intrapersonal experience and its intergroup consequences of failing to confront bias.

In our first research (Chapter 2), in order to contribute knowledge to this gap in the literature about non-target non-confronters, we aimed to investigate how people become more prejudiced following witnessing, and not confronting prejudice and discrimination. We suggest that this phenomenon occurs as product of self-justifying dissonance-reduction strategy whereby people aim to reconcile and justify their inaction by changing their (outgroup) attitudes. In other words, people experience the intrapersonal costs of not confronting, i.e., psychological discomfort, which then generates an intergroup cost in terms of amplified prejudice. In our second research (Chapter 3), we aimed to investigate ways to translate *anticipated* intrapersonal cost into an intergroup benefit. Specifically, we tested how (anticipated) personal moral cost would motivate people to confront prejudice. In this research, we provide insight into a mechanism whereby if a person cares about being non-prejudiced, the potential loss of one's sense of morality if action is not taken can actually trigger confronting behavior. These research projects will be shortly overviewed in the next section, and each research comprised the next chapters of this dissertation.

Overview of the present research

In my dissertation research, I investigated the social psychology of witnessing and (not) confronting expressed prejudice and discrimination (*hereafter* prejudice). Specifically, I tested (1) the self-justifying harmful consequences of bystanders' inaction on their own intergroup attitudes, and (2) potential moral messages that could promote bystanders' speaking up in face of prejudice. In the present research, I focused on a „bystander", who has an opportunity to confront the source of prejudice, and who is not a member of the stigmatized group that is targeted by the witnessed prejudice. We conducted our experiments in two countries, in the United States and in Hungary, across various intergroup contexts, where the outgroup is either a racial, ethnic or religious minority. For the purpose of the current research, to test actual confronting, I developed an online behavioral paradigm, where participants witness a prejudicial slur about an outgroup and discriminatory act against an outgroup individual and have the opportunity to confront the perpetrator. (This paradigm is used across all studies except for Study 4.)

The goal of our first research (reported in Chapter 2) is to identify a harmful consequence of not confronting prejudice through examining its impact on bystanders' own prejudicial attitudes. We draw on cognitive dissonance and self-justification theories to propose and test that people who witness prejudice and do not contest it (albeit having an opportunity to), subsequently endorse more negative outgroup attitudes and trivialize the witnessed incident – all in order to justify and reconcile their attitudes with their inaction. We also aimed to indirectly measure this dissonance-induced self-justification process through different methods (control groups, outcome measures, boundary conditions). We conducted five online experiments across two countries (N = 922), in the US and in Hungary, in intergroup contexts where the outgroup minority in the US was either African American (pilot studies), Muslim American (Study 2), or Latinos (Study 3), or Jewish in Hungary (Study 1). Across all studies in this research, we used the online behavioral witnessing paradigm. In Studies 1–3, we used a mixed within- and between-subjects design, where we assessed participants both prior and following witnessing the prejudiced (or control) event (pre- and post-test). This design enabled us to test overtime changes in prejudice among those who did not confront, and to compare those changes to control groups. Results confirmed our predictions. We found that those who did not confront prejudice *became* more prejudiced compared to their initial attitudes (studies 1-3). Moreover, following the incident, non-confronters' were more prejudiced and trivialized the incident more than those who did not witness any bias (pilot studies), and those who witnessed the same prejudice but had no opportunity to confront (pilot study and Study 3), and those who did not confront different, non-intergroup type of bias (studies 1–3). Supporting our proposed theoretical mechanism, this effect was not true for those who did not initially value confronting prejudice, and thus needed no justification for not confronting.

Based on the findings of the first research, we developed our second research (reported in Chapter 3), to identify moral messages about *prospective* intrapersonal costs that may motivate confronting and can be utilized as a potentially effective intervention tool. In this work, across two online experiments (N = 707) conducted in the US, we investigated how the prospect of moral loss (failure) or gain (success) in relation to intervening can motivate people to confront prejudice, depending on people's initial moral commitment to non-prejudice. Drawing on research on regulatory focus and prospect theory, we predicted and tested that a moral loss framing/mindset would significantly increase confronting tendencies among those strongly morally committed to non-prejudice (possibly to safeguard their moral self-concept), but not among those weakly committed. We also predicted that a moral gain framing/mindset would drive confronting among those who are weakly committed to non-prejudice (possibly to enhance their moral self-concept) and would not affect those strongly committed. We conducted our studies in the US in the intergroup context with Latino and Muslim outgroup. In Study 4, participants were presented with prejudiced (vignette) scenarios, and we varied the framing of moral considerations involved (loss vs. gain vs. control) and assessed (self-report) confronting intentions. In Study 5, participants went through an online moral mindset intervention that we designed. After a few days, we tested their actual confronting with using our behavioral paradigm. We found partial evidence to our predictions. Across studies, as predicted, a loss framing/mindset led to more confronting (compared to the control condition) among those highly committed to non-prejudice. Opposed to prediction, confronting in the gain condition was not significantly different than in the control condition at any level of moral commitment to non-prejudice.

Chapter 2: The Motivated Prejudice Effect – Endorsing Negative Intergroup Attitudes to Justify Not Confronting Prejudice

This chapter is based on:

Szekeres, H., Halperin, E., Kende, A., & Saguy, T. Endorsing Negative Intergroup Attitudes to Justify Not Confronting Prejudice. (under review).

Introduction

Historical and empirical precedent suggests that people often fail to stand up against prejudice and discrimination. The apparent negative consequence of not confronting in such situations is the failure to challenge the perpetrator's actions. Indeed, confronting prejudice can be effective in changing perpetrators' beliefs and reduce prejudice (e.g., Burns & Monteith, 2019; Chaney & Sanchez, 2018; Czopp et al., 2006). We here look beyond the perpetrator and investigate the impact of not confronting on the bystanders' beliefs. This focus enables us to examine a destructive trend whereby not confronting prejudice and discrimination against an outgroup changes the non-confronter's own attitudes about the outgroup – for the worse. Specifically, we propose that observers of prejudice who are given an opportunity, yet do not confront, would subsequently endorse more negative intergroup attitudes, in order to justify and reconcile with their prior inaction. By studying such a motivated prejudice process, we can identify an understudied route via which prejudice and discrimination perpetuate and intensify over time.

On many occasions, people may feel upset when witnessing prejudice and discrimination (e.g., Schmader et al., 2012; Torres et al., 2019), but nevertheless may not act against it. For example, heterosexual participants who imagined witnessing a homophobic slur reported higher intentions of confronting than people who actually witnessed the slur (Crosby & Wilson, 2015). Similarly, even though White Americans anticipated taking action against someone who expressed racism, those put in that actual situation forewent punishing the perpetrator (Karmali et al., 2017; Kawakami et al., 2009).

The inconsistency between beliefs about how one should react to prejudice and one's actual (non-confronting) behavior can give rise to psychological discomfort. Women who generally believed they should confront sexism and were made to think of instances in which they failed to do so, experienced guilt, regret and obsessive rumination (Shelton et al., 2006). Similarly, women reported more dissonance (e.g., "I feel a little conflicted about how I responded") when they imagined ignoring a sexist remark compared to having no chance to confront (Mallett et al., 2019). Those who are not the target of prejudice can experience similar feelings. For example, White Americans who felt they should not behave in a prejudiced manner towards minorities but were made to consider how they might actually do so, experienced discomfort (Voils et al., 2002; Zuwerink et al., 1996).

Based on theories of cognitive consistency, when people experience such psychological discomfort, they are motivated to reduce it by employing strategies of changing one of the elements causing the internal inconsistency or dissonance, such as changing the behavior or the relevant attitude and belief that contradicts the behavior (Abelson et al.,1968; Festinger, 1957). Given that the initial counter-attitudinal behavior cannot be changed, people will alter their attitudes instead (Aronson & Carlsmith, 1963; Kerr & Kaufman-Gilliland, 1997). In the current research we apply this logic of reducing inconsistency to observers' inaction in face of prejudice.

Previous research points to the possibility that not confronting bias can trigger dissonance-induced self-justification and lead to (seemingly counterintuitive) changes in attitudes. Specifically, Rasinski and colleagues (2013) found that female participants who valued confronting (socially inappropriate behavior in general) and were given opportunity to confront a sexist remark, but did not, subsequently made more positive evaluations of the sexist person and devalued the importance of confronting, compared to when no opportunity was given for confronting. Researchers argued that this effect is driven by motivation to reduce cognitive dissonance. That is, seeing the perpetrator as "not that bad" and confronting as less important, reflected female participants' attempts to reduce inconsistency between attitudes about confronting sexism and their failure to actually do so (Rasinski et al., 2013; for similar results see Mallett et al., 2019).

Based on prior work reviewed above, in the current research, we identify a potential cycle of rising prejudice and discrimination. We tested the effect of witnessing and not confronting prejudice among those who do not belong to the targeted outgroup. Specifically, we expected that those who witness prejudice and do not confront (while given an opportunity to) would be motivated to change their attitudes about that outgroup in order to obtain consistency between their beliefs and their inaction (i.e., not confronting). In the context of not confronting, an effective coping strategy would be rendering the incident justified, namely, viewing the incident as based on reasonable judgment and having a kernel of truth in it (i.e., along the line of "after all, they are kind of like that"). Thus, we propose that not confronting prejudice and discrimination would lead to escalation of negative outgroup attitudes.

Besides attitude change, people may engage in additional (often used) dissonance-reduction strategies (for review see McGrath, 2017), which would be detrimental to intergroup relations, namely in *trivialization* (aka. minimizing the significance of the element causing dissonance; Festinger, 1957; Simon, Greenberg, & Brehm, 1995), and in *denial of responsibility* (Gosling et al., 2006). Research shows that people can engage in multiple modes of dissonance reduction, even for the same dissonant event (McGrath, 2017), therefore, besides prejudice increase, we tested trivialization and responsibility denial as other outcome variables. We predicted that people who do not confront will also be motivated to trivialize the prejudiced event, that is, reappraise it as not sufficiently serious to warrant confronting, and also motivated to deny responsibility for acting in the situation. Such intergroup attitude changes as consequences of not confronting prejudice are harmful as they build tolerance for prejudicial atrocities in the long-run which in turn likely to go uncontested.

The Present Research

We ran two pilot and three (primary) experiments to test our prediction that when people witness prejudice and do not confront, albeit given an opportunity to, they will endorse more negative outgroup attitudes and will also trivialize the prejudiced event and deny responsibility for acting. For the purpose of the current research, we developed a behavioral paradigm, in which participants observed and played an online game, where they witnessed a player being prejudiced and discriminatory against an outgroup member and had an opportunity to confront the prejudiced player.

Following, preliminary qualitative pilot study, pre-tests and pilot studies, in Studies 1–3, using a mixed within- and between-subjects design, we assessed participants both prior and following witnessing of a prejudiced event (pre- and post-test). This design enabled us to test overtime changes in attitudes among those who did not confront, and to compare those changes to control groups, in order to show that people *come* to endorse more negative outgroup attitudes as a function of witnessing and not confronting prejudice.

In studies 1–2, in the control condition, participants observed another type of prejudice not rooted in intergroup membership (but “interpersonal”) and had an opportunity to react. We predicted no attitude change for those who did not confront interpersonal bias, compared to intergroup bias. This would show that the proposed effect is not specific to a personality type who does not confront socially inappropriate behavior in general (or not about assertiveness), nor is it a derogatory response resulting from a deflated self-esteem (Fein & Spencer, 1997) that would be brought upon by any personal failure of not confronting. Similarly, to rule out an explanation that intergroup non-confronters are characteristically more conservative, non-egalitarian or prejudicial than the control groups, we tested across all studies, differences in baseline outgroup attitudes, and individual socio-political orientations (e.g., Social Dominance Orientation; this differed based on country).

In Study 3, we added another control condition, where participants observed the same intergroup prejudice but did not have an opportunity to confront – they were only exposed to prejudice (we also had this condition in pilot study 2). This allows to test the dissonance-induced self-justification account. We reasoned those participants who were not given a chance to confront would not experience dissonance and engage in attitude change, because they had external justification for staying silent (Leippe & Eisenstadt, 1999). If there is no intergroup attitude change following only exposure to prejudice that allows us to rule out desensitization, persuasion or change in normative context (e.g., Blanchard et al., 1991; 1994) or victim blaming triggered by just-world beliefs (e.g., Lerner & Simmons, 1966). To further support the dissonance account and to demonstrate boundary conditions to the proposed effect, we tested and predicted that those for whom *not* confronting prejudice do not contradict their personal values (thus have no need to justify their inaction), will not show the motivated prejudice effect.

To establish external validity, each study was run in a different intergroup context, in the US with African American (pilot studies), Muslim American (Study 2), or Latinx outgroup (Study 3), and in Hungary with Jewish outgroup (Study 1). The witnessed incident and measures were framed around intergroup trust and liking, however the actual slurs varied to fit the predominant prejudice about the target outgroup, and outgroup attitude measures varied accordingly (mostly in study 1) and also by keeping in mind what is used in the literature (mostly in study 2-3). Finally, we report how we determined sample sizes, all data exclusions, manipulations, and measures in the manuscript or in appendix, (all additional measures are reported in Appendix C). All data and analyses of Studies 1-3, and the pre-registration of Study 3 is available here: https://osf.io/36ay8/?view_only=2f41a047b78b46e99055e5255a558336.

Preliminary pilot study

Prior to our experiments we conducted an in-lab preliminary pilot study in order to gain initial insight into people's reactions to racism and to explore their thoughts and feelings about such a situation. To this end, in an Israeli international college, Caucasian students (n = 11) were invited to the lab for IQ/cognitive testing. The study was video-taped. In the lab room, the participant observed that once the previous participant (a Black confederate) finished and left, the (white) RA scrambled up his test, threw it to the trash and made an insulting remark: "We cannot rely on *THEIR* data. It will just make the average lower" (sometimes adding: "you know Blacks"). We provided around 5-10 minutes for the participant as an opportunity to confront, which was followed by an in-depth interview.

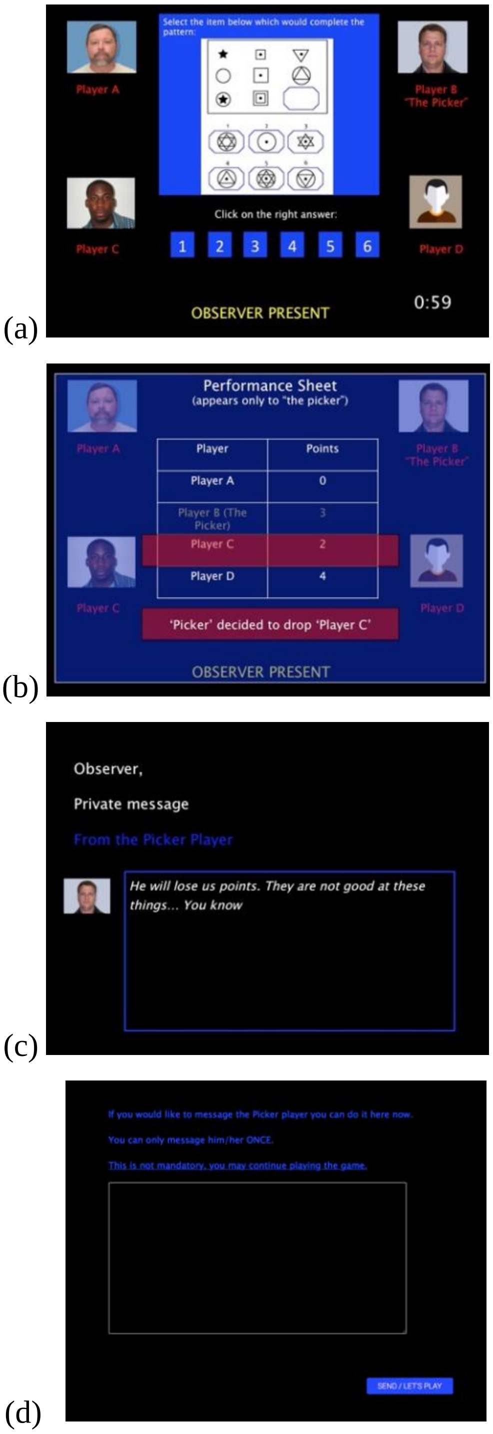
Out of the 11 students nobody confronted but many showed signs of intentions of doing so, which was evident from their facial expressions (and for example, opening their mouth to speak, but then stopping themselves). During the interview a few patterns of responses emerged. First of all, most participants did not mention (or admit) that they witnessed a racist situation only once they realized it was a set-up, and some participants needed more time to admit it (e.g., “I didn’t hear it”, “Oh I heard we can`t use *air*”). Based on our judgment, the majority of their responses fell into the following three categories: (1) Reporting negative emotions and intentions to confront (e.g., “I felt angry and frustrated” “I would have said something in the end of the study”, “The shock kept me quiet”, “I’d thought about it the whole day”); (2) Undermining the severity of the incident (e.g., “He did not cause actual harm”, “He [Black] gets his credits anyway”, “Nobody got hurt”, “Who cares, it doesn’t have any consequences”, “He wasn’t personally hurt”); and (3) Denying responsibility or controllability (e.g., “It’s none of my business”, “I cannot change his opinion anyway”, “It’s strange, it wasn’t my business, I was just relaxing here, it’s my personality, I don’t get involved.”, “I was afraid I won’t get my credits”). An additional two categories emerged although less intensively: (4) Empathy towards the racist confederate (e.g., “I just didn’t want to be rude”, “He looked busy … Then we would get into an argument, and it’s not nice to take his time”); and (5) Victim blaming (e.g., “I thought he threw it away because his test just wasn’t good”; “… maybe his attire, and I look more sophisticated, so maybe he based it on surface.”)

Based on participants’ facial expressions in response to the racist remark and based on the discussion during the interviews, we assessed that the in-lab scenario was emotionally intrusive for the participants and due to corresponding ethical considerations and to ease the feasibility of conducting multiple studies, we decided to place our paradigm online. Overall, the in-lab study gave us valuable insight and ideas for the development of the online paradigm and how to further progress.

Pilot studies: U.S. with African American outgroup

Given the novelty of the paradigm and research inquiry, we aimed to perform an initial test of our hypothesis and conducted two pilot experiments (N=71 and N=183). White American participants observed an online game, we designed and pre-tested, called “Logic-IQ game”, where players answer logical questions (see Figure 3 for scenes). Participants either witnessed a white player unfairly eliminating a Black player from the game and then privately messaging the participant with a prejudiced remark (about intellectual abilities), or witness a (white) player being eliminated with neutral message (control condition). In both conditions, participants had an opportunity to reply to the player’s message (aka. confront). In pilot study 2, we had an additional control condition where participants witnessed the (same) prejudice but had no opportunity to reply to the message (exposure condition, similarly to Study 3). We tested and found that participants who had a chance but did not confront prejudice, had subsequently more negative outgroup attitudes (less willing to support a Black education program) and trivialized the intergroup prejudiced incident more (only in pilot study 2), compared to control condition(s).

Figure 3. Scenes from the Logic-IQ game: (a) During a question posed to players; (b) Performance sheet with players' earned points and showing that Black player is eliminated by the prejudiced (Picker) player; (c) Picker player's prejudiced message; (d) Message box providing an opportunity to respond to the prejudiced (Picker) player. (Pictures taken from the Chicago Face Database; Ma, Correll, & Wittenbrink, 2015)



(a)

Select the item below which would complete the pattern:

Click on the right answer:

(b)

Performance Sheet
(appears only to "the picker")

Player	Points
Player A	0
Player B (The Picker)	3
Player C	2
Player D	4

'Picker' decided to drop 'Player C'

(c)

Observer,

Private message

From the Picker Player

He will lose us points. They are not good at these things... You know

(d)

If you would like to message the Picker player you can do it here now.

You can only message him/her ONCE.

This is not mandatory, you may continue playing the game.

Due to our research predictions regarding those who choose not to confront, we excluded participants who confronted prejudice (see Rasinski et al., 2013 for a similar approach), and compared non-confronters to the control group(s). To test that the found effects are not explained by individual-level differences, we measured and found no baseline differences between these groups on socio-political–intergroup orientations (same measures as in Studies 2-3), and results remained significant when controlling for these variables. However, those who tend not to confront may still have particular characteristics that set them apart from participants in the control groups – and thus it is not only the manipulation that may drive the effects. While the pilot studies provided initial support for our predictions, due to this limitation, we report them as supplementary information (see Appendix B). To overcome the selection issue, which is inherent to our research question, in the next studies we used pre-posttesting. This design enabled us to test overtime changes within non-confronting participants and to compare those changes to control group(s).

Study 1: Hungary with Jewish outgroup

In study 1, we used mixed within-between-subjects design to test our prediction of negative escalation of intergroup attitudes following not confronting prejudice. To rule out the derogatory, and non-confronting personality account, we used an interpersonal prejudice control condition. We conducted the study in Hungary with Jewish outgroup. Antisemitism has been a problem facing Hungarian society in the past and present (Kovács, 2014). There is a predominant prejudice in Hungary about Jews being manipulative and untrustworthy (Kende, Nyúl, & Hadarics, 2018; Kovács, 2014). In order to correspond to the intergroup context, participants went through an online behavioral paradigm we designed based on a “Trust Game”, witnessed a prejudicial slur about not trusting Jews with money, and then responded to questions about perceived trustworthiness of Jews (outgroup attitude measure).

In the study, participants first responded to a pre-survey, which included pre-test assessment of perceived trustworthiness. Additionally, to assess socio-political–intergroup orientations we measured their political ideology (conservative–liberal, right–left-wing) and agreement with a political antisemitism scale (from sociology research; Kovács, 2014).[20] A few weeks later participants returned to post-test, where they played the Trust Game. They were randomly assigned to intergroup or interpersonal prejudice condition. In the intergroup condition, participants observed a player discriminating and making a prejudiced remark about a Jewish player. In the interpersonal condition, participants observed a player mistreating another player based on simply disliking his name. All participants had an opportunity to respond to the perpetrator (i.e., confront). Following the game, in a seemingly unrelated survey among filler scales participants responded again to outgroup attitude measure (post-test assessment). At the end of the experimental session, across all conditions, we assessed (cunningly) the trivialization of the *intergroup* prejudiced event, and the extent they deny responsibility for intervening (note that it was not possible to test trivialization and responsibility at pre-test as it referred to the prejudiced incident).

Method

Participants and procedure. For summary of all study samples see Table 1. A questionnaire on various social issues was sent out at by multiple research groups in a university and 530 Hungarian students from various schools/departments completed it for course credits in the spring of 2017. The sample size was determined by the number of eligible students in the subject pool (for further details see Appendix D). We included the *perceived trustworthiness of Jews scale,* in which respondents indicate the extent qualities listed are typical of an average Jewish person (on a scale from 1 ‘not typical at all’ to 7 ‘completely typical’): manipulative (reverse), insincere (reverse), trustworthy (α = .78; list

[20] Antisemitism in Hungary is known to be strongly connected to political affiliation (Kende et al., 2018; Kovács, 2014).

included filler items). The survey also included questions of basic demographic (age, gender, SES, education level; see Appendix A for demographic questions across studies) and political ideology (conservative–liberal, right-wing–left-wing; 1–7). Participants also responded to the political antisemitism scale (Kovács, 2014) that included items like “There is a secret Jewish conspiracy that determines political and economic processes” (7-point scale; 6 items; α=.93; see Appendix A for socio-political–intergroup orientation measures across studies).

Five weeks later, we approached the students again, who had no information that the surveys are connected. Participants were told we are testing how observing, and gender of players/observers influence trusting behavior. Around 35% of students returned to participate, and the 190 participants (77.7% female, M_{age}=20.47, SD_{age}=1.57; we had two attention checks[21] but all participants who returned answered them correctly) were randomly assigned to intergroup (n=97) or interpersonal bias conditions (n=93). Participants played the Trust Game and then were directed to an allegedly independent survey on social issues, where they again responded to the *perceived trustworthiness scale* (3 items, α=.75).

At the end of the study, in order to assess *trivialization* and *responsibility denial* regarding the intergroup prejudiced event, we told participants that a survey respondent reported about a possibly prejudiced player. In the intergroup condition, we were vague in this description (in order to decrease suspicion) and asked participants if they encountered such behavior; while in the interpersonal condition we described the prejudiced situation exactly as it occurred in the intergroup condition (for exact description see Appendix A). Then participants received the trivialization and responsibility denial scales (both referring to the intergroup prejudiced event in all conditions), for which we adapted the emergency and responsibility subscales of the Confronting Prejudiced Responses measure (CPR; Ashburn-Nardo, Morris, & Goodwin, 2008).

[21] Among statements, we included items “This is an attention check question. For response mark the mid-point answer.” And “This is again an attention check question. Please mark the strongly agree response.”

For *trivialization*, we asked participants about the statement and behavior of the [prejudiced] player (on a 7-point dis/agreement scale): (1) The behavior requires an immediate response. (2) The behavior hurt other people. (3) Something should be done right away to stop the behavior. Items were reversed and averaged to a trivialization scale (α=.78). For *responsibility*, we included the following statements: (1) I felt/would feel personally responsible for doing something about the behavior. (2) It was/would not be my place to say or do something. Items were reversed and averaged to a responsibility denial scale (α= . 58, *r*=.41, *p*<.001).[22] Finally, participants were debriefed (Appendix A). All study materials were in Hungarian. See Figure 4 for study procedure across studies.

Stimuli and confronting. We based the "Trust Game" on the behavioral economic game (e.g., Berg et al., 1995), where Player A decides how much money of an initial endowment to send to Player B. The sent amount is then multiplied by some number and Player B decides how much to send back to Player A. Participants were given instructions and quiz on the game (Appendix A) and were told they will observe and then play the game. We emphasized for participants that in this game, the most beneficial behavior is sharing more money with the opponent.

Participants entered a seemingly different online surface to observe the game, which was pre-programmed (players appeared with names). They were asked for and appeared with their nickname throughout the game to make them feel present in the situation. All participants were assigned to observe a player called Márk (average Hungarian name), and then observed two decoy rounds and exchanged (programmed) messages with Márk. Then, participants in the intergroup condition were presented with Slomó (Jewish name) as the new opponent to Márk, while participants in the interpersonal condition were presented with 'Zsolt' (average Hungarian name). In all conditions, they saw that Márk chooses

[22] We had an additional item in this scale "It is someone else's responsibility for doing something against this behavior (e.g., the researcher)" (in Study 1), and "I would expect someone else to take responsibility for doing something." (in Study 2 and 3). As a post-hoc decision, due to low internal consistencies across studies when including this item (α=.37 in Study 2, α=.45 in Study 3), we excluded it from analyses.

to give no money to Slomó/Zsolt (but keep it all to himself) – which is unlike his behavior in the previous rounds. Then, Márk privately messages the participant saying “well, I won’t trust these rothschilds with money” (intergroup)/ “I have a bad feeling about people named Zsolt” (interpersonal). (For scenes from the English version of the game in Study 2 see Figure 5).

Under the prejudiced message (both conditions), participants could either press ‘reply’ or ‘continue game’. Those who replied received a notification that the message was read (but received no response from Márk), and the game continued. In the end, we allowed participants to play the game themselves. Those who continued and those who replied in a non-confronting manner were coded as non-confronters and we continued data analyses with them (for a similar approach see Rasinski et al., 2013). The rest of the responses were coded as either expressing confrontation, [23] agreement with the perpetrator, suspicion (thus weeding out those questioning the realness of the manipulation),[24] or whether it had unclear meaning.

[23] For analyses with confronters across studies see Appendix D.

[24] We also checked suspicion in the an open-ended question at the end of study (“Please feel free to leave any comment/s or remark/s you may have.”) but participants did not suggest that they knew that the game was fabricated and/or that the racist remark was the actual aim of the study.

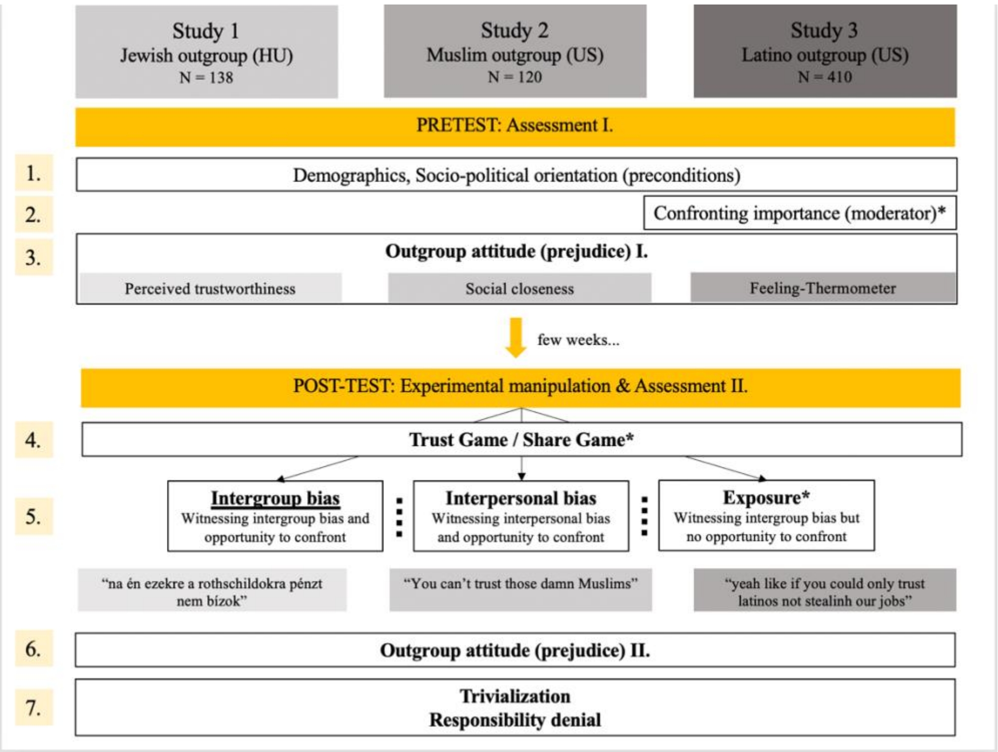
Table 1. Summary of all study samples.

	Study 1	**Study 2**	**Study 3**
country	Hungary	U.S.	U.S.
recruitment	university students from all disciplines	mTurk (American participants)	mTurk (American participants)
compensation	course credits	monetary	monetary
sample size at pre-test	530	300	630[25]
sample size at post-test prior to exclusion	190 (n=97 in intergroup, n=93 in interpersonal)	206 (n=105 in intergroup, n=101 in interpersonal)	485 (n=163 in intergroup, n=165 in interpersonal, n=157 in exposure)
exclusion criteria	(1) 2 attention checks (n=0 failed) (2) *not* non-confronting response (n=22 intergroup, n=30 interpersonal) (intergroup confronting rate was 8%) [a]	(1) Only those who did not identify as Arab/Muslim were invited to post-test (n=3 excluded). (2) 2 attention checks (n=10 failed) (3) *not* non-confronting response (n=52 intergroup, n=24 interpersonal) (intergroup confronting rate was 42%) [a]	(1) Only those who identified as White/Caucasian/European American, and those who passed the 1 bot check question were able to fill out the pre-survey. (2) 1 attention check at post-test (n=14 failed) (3) *not* non-confronting response (n=44 intergroup, n=17 interpersonal) (intergroup confronting rate was 24%) [a]
final sample size after exclusions	138 (n=75 in intergroup, n=63 in interpersonal)	120 (n=49 in intergroup, n=71 in interpersonal)	410 (n=112 in intergroup, n=142 in interpersonal, n=156 exposure)
sensitivity test (power) b	Reliably detecting effects from η^2_p = .01	η^2_p = .02	η^2_p = .02
gender demographic	76.8% female, 21.7% male, 1.4% don't wish to answer.	52.5% female, 47.5% male	49.8% female, 49.8% male, 0.5% other
age demographic	M_{age}=20.57, SD_{age}=1.57, range: 18–26	M_{age}=36.52, SD_{age}=12.73, range: 18–74	M_{age}=40.98, SD_{age}=13.19

Note. [a] For details on responses see SM. [b] For detailed sensitivity power analyses see SM.

[25] Aimed for 631, but there was a bug in the system.

Figure 4. Study procedure across studies.


Study 1
Jewish outgroup (HU)
N = 138
Study 2
Muslim outgroup (US)
N = 120
Study 3
Latino outgroup (US)
N = 410
PRETEST: Assessment I.
1.
Demographics, Socio-political orientation (preconditions)
2.
Confronting importance (moderator)*
3.
Outgroup attitude (prejudice) I.
Perceived trustworthiness
Social closeness
Feeling-Thermometer
few weeks...
POST-TEST: Experimental manipulation & Assessment II.
4.
Trust Game / Share Game*
5.
Intergroup bias
Witnessing intergroup bias and opportunity to confront
Interpersonal bias
Witnessing interpersonal bias and opportunity to confront
Exposure*
Witnessing intergroup bias but no opportunity to confront
“na én ezekre a rothschildokra pénzt nem bízok”
“You can’t trust those damn Muslims”
“yeah like if you could only trust latinos not stealinh our jobs”
6.
Outgroup attitude (prejudice) II.
7.
Trivialization
Responsibility denial

Results

First, we coded responses to the prejudiced message. Among participants in the intergroup condition (n=97), 77% (n=75) did not confront the remark (n=48 continued the game without reply; n=27 replied in a non-confronting manner, e.g., “Let’s play”), constituting the intergroup non-confronting group (see Rasinski et al., 2013 for a similar approach). The rest of the responses were either confronting (8%, n=8; e.g., “This is discrimination”), or expressed prejudicial agreement to the remark (n=11, “yeah I don’t like Jews either”), or questioned the realness of the scenario (n=3). In the interpersonal condition (n=93), 68% (n=63) did not confront the remark (n=43 continued the game; n=20 had non-confronting reply), constituting the interpersonal non-confronting group. Others confronted (24%, n=22), expressed agreement (also not liking people with that name; n=4), or suspicion (n=4). We continued the analyses with those who did not confront (n=75 and n=63). For details on sensitivity power analyses across studies see Appendix D.

Results from the pre-survey revealed no significant *baseline* differences between intergroup non-confronters and interpersonal non-confronters on demographic, political orientation (right-left, conservative-liberal), political antisemitism, or on perceived trustworthiness of Jews (all *p's*>.25).

Next, we ran a mixed ANOVA and obtained significant interaction between time (pre-test vs. post-test) and condition (intergroup non-confronting vs. interpersonal non-confronting) on perceived trustworthiness of Jews, see Figure 6. According to our main prediction, repeated measures (simple effects) analysis showed that the intergroup non-confronting group reported significantly less perceived trustworthiness at post-test compared to pre-test. No significant change occurred among interpersonal non-confronters. For means, standard deviations and inter-item correlations across studies see Table 2, and for statistical values across studies see Table 3.

To test differences *between groups* following the game while taking into consideration the pre-test measurement, i.e., baseline differences, we ran a between-subjects Multivariate ANCOVA on perceived trustworthiness (post-test), on trivialization and on responsibility denial – while controlling for trustworthiness scores obtained in the pre-survey ("pre-scores" hereafter). As predicted, we found that following the incident, intergroup non-confronters, compared to interpersonal non-confronters reported significantly less *perceived trustworthiness,* more *trivialization* (also significant when not controlling for pre-scores, instead conducting independent samples t-test, *p*=.004), and more *denial of responsibility* (also significant without covariate, *p*=.041). See Figure 7. Note that repeated measures analysis is not available for trivialization nor responsibility (because it referred to the prejudiced event, so it was measured only at post-test).

Discussion

The findings of Study 1 supported our predictions by demonstrating negative overtime outgroup attitude change (only) among those who witnessed but did not confront intergroup prejudice. Trivialization of the intergroup prejudiced event and responsibility denial was also significantly higher in the intergroup non-confronting than in the interpersonal non-confronting group (also while controlling for baseline prejudice). The lack of significant change among interpersonal non-confronters suggest that the found effect is not related to a certain personality type who does not confront in general, nor to experiencing personal (moral) failure to confront just any type of negative treatment. We also observed that prior to witnessing prejudice, there were no baseline differences between non-confronting groups on outgroup attitudes (nor on demographic and socio-political–intergroup orientations). This further supports that the observed attitude change is unique to witnessing, and not confronting intergroup prejudice. In the next study, we aimed to further establish the motivated prejudice effect in a different cultural and intergroup context.

Study 2: U.S. with Muslim outgroup

In study 2, we aimed to conceptually replicate the effects revealed in study 1 and increase the external validity of our findings. We placed this study in a different cultural context, in the US with American Muslims as the outgroup. Besides context-relevant differences, study design, procedures and predictions were like in Study 1. Materials reflected the predominant prejudice of fear and distrust towards Muslims (e.g., Oswald, 2005). All participants played the Trust Game and witnessed either intergroup or interpersonal prejudice and had an opportunity to confront. Our main outcome measure was a (willingness to) social closeness scale towards American Muslims (e.g., Oswald, 2005). To test socio-political–intergroup orientations, we measured political orientation (conservative-liberal, democrat vs. republican, Clinton vs. Trump voting), System Justification (justify the system as legitimate and necessary; SJ, hereafter; Kay & Jost, 2003), Social Dominance Orientation (preference for societal inequality; SDO; Pratto et al., 2013), Internal Motivation to Respond without Prejudice (internalized ideals of oneself as non-prejudiced person; IMS; Plant & Devine, 1998).

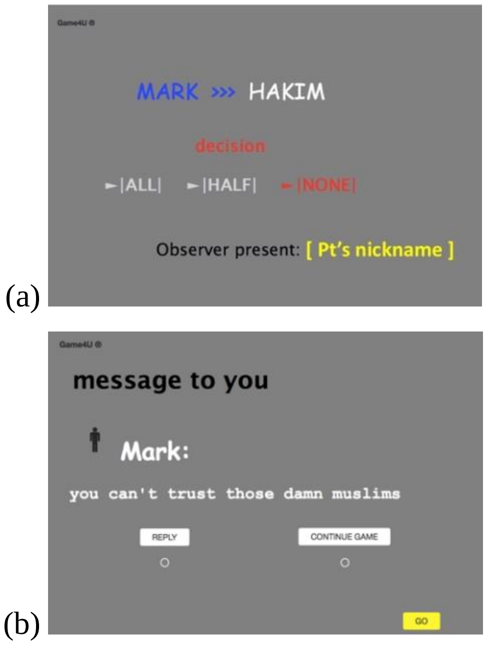
Method

We recruited 300 U.S. residents through Amazon's Mechanical Turk /Turk Prime, who filled out our pre-survey for monetary compensation (50 cents) in the spring of 2017. For how sample size was determined see Appendix D. The pre-survey, among filler scales, composed of demographics (age, gender, education, SES), socio-political–intergroup orientations (α=.88 for SJ, α=.84 for SDO, α = .95 for IMS), and the social closeness scale (adapted from Oswald, 2005). This scale asked about Muslims living in the US and began with the stem "How would you feel about" followed by 8 items (α=.97) on a 6-point Likert scale from 1=*I feel very unfavorably* to 6=*I feel very favorably*: "living next door to Muslims?", "becoming a close personal friend to Muslims?", "your son or daughter marrying a Muslim?", "inviting Muslims to your home for dinner?", "having a Muslim roommate?", "working closely with Muslims on a job?", "if your child were in the same class as Muslims?", and "trusting a Muslim to look after your child?".

Five weeks later the same respondents received a message advertising the study (except those identified as Arab/Muslim in pre-survey, n=3). Participants did not know the surveys are connected. With a 33% dropout rate, 199 respondents returned and completed the study. They were randomly assigned either to intergroup or interpersonal prejudice conditions. Participants who failed the attention checks (n=3) were excluded from analyses, leaving 196 participants (n=101 in intergroup and n=95 in interpersonal; 50.3% female, M_{age}=36.23, SD_{age}=12.27).

The experimental procedure was similar to Study 1. Participants observed and played the Trust Game (See Appendix B for a pilot test for the paradigm in the US). Participants in the intergroup condition witnessed a player (Mark) discriminating against a Muslim player (Hakim) by not sharing any money with him (see Figure 5) and then messaging the participant with an explicitly prejudiced remark about Muslims: "You can't trust those damn Muslims" (see Figure 5b). Participants in the interpersonal condition also observed the same player not sharing money with his opponent and then messaging the participant "I have a bad feeling about people named Jeff". Confronting was measured and coded like in Study 1. Following the game, participants were directed to an allegedly independent survey on social issues, where they again completed the *social closeness scale*. In the last section like in Study 1, they responded to the *trivialization scale* (3 items, α=.91) and *responsibility denial* scale (2 items, α=.57, r=.39, p<.001).

Figure 5. Scenes from the game: (a) Prejudiced player (Mark) is playing with the Muslim player (Hakim) and denies him money; (b) Mark messages the participant with a prejudiced remark about Muslims.



you can't trust those damn muslims

Results

First, we coded responses to the prejudiced message. Among participants in the intergroup condition (n=101), 49% (n=49) did not confront (n=44 continued the game without reply; n=5 pressed reply but left the textbox empty, or gave non-confronting response e.g., "He may have been Hindi"). Rest of the responses were either confronting (42%, n=42; e.g., "That sort of thinking is disgusting and pure racism."), or expressed prejudicial agreement to the remark (n=1, "Nope you sure can't"), or questioned the realness of the scenario (n=2), or agreed to the prejudiced remark (n=1), the response was unidentifiable (as to whether it was confronting or not, agreeing or not, or figured out the study; n=7). In the interpersonal prejudice condition (n=95), 79% (n = 71) did not confront the remark (n= 43 continued the game; n=27 had non-confronting reply), and the rest agreed (n=2) or confronted (21%, n=20), or was unclear (n=2). We continued the analyses with non-confronters (n=49 and n=71).

Preliminary analyses revealed no significant baseline differences (pre-survey) between intergroup non-confronters vs. interpersonal non-confronters on demographic (*p's*>.25), socio-political–intergroup orientation (*p*'s>.10), and on social closeness to Muslims (*p*>.25).

Like in Study 1, we found significant interaction between time (pre vs. post-test) and group (intergroup non-confronting vs. interpersonal non-confronting) on social closeness (see Figure 6). As predicted, repeated measures (simple effects) analysis revealed that participants in the intergroup non-confronting group expressed significantly less social closeness after compared to prior to the prejudiced event. No attitudinal change occurred for participants in the interpersonal non-confronting group.

Like in study 1, to test differences *between groups* following the game while considering baseline attitudes, i.e., controlling for social closeness pre-scores, we conducted Multivariate ANCOVA. As predicted, we found that after the incident intergroup non-confronters, compared to interpersonal non-confronters reported significantly less *social closeness,* and more *trivialization* (also significant without covariate, *p*=.010). Contrary to prediction, there was no significant difference on *denial of responsibility* (also significant without covariate, *p*=.148). See Figure 7.

Discussion

The findings of Study 2 further support our prediction that not confronting intergroup prejudice, when given an opportunity to, results in escalation of negative outgroup attitudes, and in trivialization of the witnessed prejudice. Thus, we found the proposed effect in a different country and intergroup context than in Study 1, which contributes to the generalizability of our findings. However, unlike in study 1, we did not find a significant effect on responsibility denial, which somehow may be due to the different intergroup context, or what we found in study 1 was a false positive result. At this point we do not have a valid explanation for this null-effect.

A potential limitation of studies 1 and 2 is the use of the interpersonal prejudice manipulation as a control condition. A remark and discrimination based on someone's name might not be as aggravating as based on being a minority, which can potentially create a possible confound. Although in both studies, confronting was similar or even higher in the interpersonal than in intergroup condition, which indicates the possibility that they were both aggravating, it is also possible that the confronting was triggered by perceiving the interpersonal slur as so surprising and "silly". To address this limitation, in the next study, a group of participants witnessed exactly the same intergroup prejudice but had no opportunity to confront.

Another limitation of Study 2, compared to Study 1, is the high level of confronting rates (42%), which is not ideal in terms of selection bias. This was likely due to the remark being more blatant and hurtful, and due to the specific climate – during an anti-Muslim wave in the US originating from the President's political campaign (e.g., Muslim-ban). Participants likely felt more inclined to respond to the remark. We found no initial individual differences between the groups on outgroup attitudes (nor on demographic or socio-political–intergroup orientations), nevertheless, in the next study, we made pro-active measures in order to decrease confronting rate.

Study 3: U.S. with Latinx outgroup

In study 3, we aimed to rule out an alternative explanation that the found effect is due to persuasion or change in the normative context, and instead to find further support for the dissonance-reduction account. In this study, we manipulated prejudice against a Latino player (US). Like in study 1 and 2, we used a mixed design, participants filled out a presurvey then returned to an experiment and either witnessed intergroup prejudice and had an opportunity to confront (intergroup condition) or witnessed interpersonal bias and had an opportunity to confront (interpersonal condition). In a third condition, the exact same intergroup prejudiced incident was witnessed, but participants had no opportunity to reply to the message (exposure condition). This control condition allowed to test the dissonance account, because participants who were not given a chance to confront are not expected to experience dissonance and engage in attitude change, because they had external justification (and no personal responsibility) for staying silent (Leippe & Eisenstadt, 1999).

Therefore, we predict that only intergroup non-confronting group would show significant overtime change in their attitudes and not the control groups, i.e., exposure and interpersonal non-confronting groups. We also predict that following the incident intergroup non-confronting group would significantly differ from both control conditions. In order to further show that mere exposure to prejudice does not drive the hypothesized effect, we also predict that the exposure condition would not significantly differ from the interpersonal non-confronting group.

To further substantiate the dissonance account, we tested the boundary conditions for the predicted effect. Our theorizing relies on the assumption that people would feel discomfort for not standing up against prejudice. This would not describe those for whom not confronting prejudice do not contradict their personal values. To this end, similar to prior research (Rasinski et al., 2013), we measured perceived importance of confronting, adapted this scale specifically to confronting prejudice, and used it as a moderating variable. We predicted that the worsening of attitudes would be less pronounced among those lower on confronting importance because for them, justification following inaction is not required as it does not contradict their personal values. (Also, we measure prior difference on importance of confronting for assurance that there are no baseline differences on this attitude.) Additionally, after the game we measured negative affect and psychological discomfort. We predicted that not confronting intergroup prejudice (vs. controls) will lead to negative affective reaction which would predict negative intergroup attitudes.

We used similar game paradigm as in Study 1-2, but slightly altered in order to flatten the confronting rate. For this reason: (1) In the instructions, we specifically asked participants to respond only when they must, to minimize conversation. (2) We increased the cost of confronting by telling participants that they will play with the player they observe (the perpetrator), thus creating a risk that if they confront him, he will later punish them by not sharing money. We called this game the "Share Game" due to these changes and also to correspond to predominant prejudice of Latinos about paying taxes and abusing social welfare (Valentino et al., 2013), accordingly we also changed the prejudicial remark.

For outgroup attitude measures, we used the Modern Racism Scale towards Latinos (Abad-Merino et al., 2013), however it is often used as a moderator (not outcome variable), and items are quite obvious and could elicit demand in the study (e.g., "Discrimination against Latinos is no longer a problem in the U.S."). Therefore, we also added another measure, which perhaps corresponds less directly with the prejudiced remark, but has been used to assess attitudes towards Latinos, the feeling-thermometer (e.g., Valentino et al., 2013; Huo et al., 2018). Measure of trivialization, responsibility denial and socio-political–intergroup orientation was the same as in Study 2. This study was pre-registered: https://osf.io/36ay8/?view_only=2f41a047b78b46e99055e5255a558336.

Method

Participants and procedure. The experimental procedure was similar to Study 2. We recruited 631 U.S. residents to fill out our pre-survey (for 50 cents; ended up with 630 workers due to a bug) in the summer of 2020. Only Whites/Caucasians/European Americans and those who passed the 1 bot check question were able to fill out the pre-survey. For how sample size was determined see Appendix D.

The pre-survey, same as in Study 2, among filler scales, included demographics and socio-political–intergroup orientations (α=.90 for SJ, α=.86 for SDO, α=.89 for IMS). Additionally, the measure of confronting importance, Feeling-thermometer, and Modern Racism Scale ("MRS"; see Appendix A). We used the *perceived importance of confronting scale* (we adapted specifically to prejudice from Rasinski et al., 2013) and asked how important are the following behaviors to them (on a 10-point scale from 1=not at all important to 10=completely important; 5 items, α=.96): How important is the issue of confronting prejudice in your opinion?; How important is it that you express disapproval with someone else's prejudicial opinions?; How important is it that you speak your mind when someone is acting in a prejudiced manner?; Do you feel that it is necessary to voice your concerns over other people's prejudicial actions?; How significant would it be if you did nothing while someone else was acting in a prejudiced manner?. These items referred to prejudice in general, and they were not specific to Latinx to make sure participants do not anchor themselves for confronting by responding to this scale and that they do not figure out the studies are connected.

For *feeling-thermometer,* we asked participants "How warm (favorable) or cold (unfavorable) do you feel towards each of the following groups" on a slider of 0 (very cold) to 100 (very warm): Asians, African American, Latinos, and White/Caucasians. (Participants were told survey questions refer to [groups in] American society).

One week later, from a different mTurk account, we returned to participants, who passed the attention check (pre-survey). Data was collected for weeks until reaching 485 participants, who returned to participate in the experiment. Those who did not pass the attention check question (post-survey), were excluded from data analyses (n=14). Remaining 471 participants (50.5% female, 49.3% male, 0.2% other, M_{age}=40.98, SD_{age}=13.19) were randomized into prejudice (n=156), exposure (n=156), and interpersonal conditions (n=159). Following the Share Game, participants responded to the negative affect (subscale of PANAS; Watson et al., 1988; α = .88) and psychological discomfort scales (Elliot & Devine, 1994; α = .84), see Appendix A. Later in a seemingly other survey, they again received the feeling-thermometer and MRS, and finally trivialization (3 items, α=.88) and responsibility denial (2 items, α=.69, *r*=.53, *p*<.011).

Stimuli and confronting. The game and procedure were similar to study 1-2, however certain changes were made. For full instructions to the game, see Appendix A. We made the game an ultimatum game (there is a giver and a receiver), instead of a trust game (they both give money). This was done so participants cannot decide to "confront" by punishing the perpetrator with not returning money to him when they play together – which would make the measurement of confronting complicated. However, we still had to make sharing the rational behavior therefore we explained that the same players usually return, and the roles can change: "Players can become a receiver to someone they either shared with or did not share (as a giver), therefore as a rule of thumb, players should always share."

In the intergroup prejudice condition, a player (Mark) discriminated against a Latino player (Sanchez) by not sharing money with him and then messaging the participant with the prejudiced remark: "yeah like if you could only trust latinos not stealinh our jobs" (there was a typo on purpose, to make it look more believable). Participants in the exposure condition witnessed exactly the same scenario but unlike in the intergroup prejudice condition, they had no opportunity to reply to Mark. Like in Study 1-2, participants in the interpersonal condition also observed the same player not sharing money with his opponent (Jeff) and messaged: "I have a bad gut feeling about people named Jeff". Confronting was measured and coded like in Study 1-2.

Results

In the intergroup condition (n=156), 72% (n=112, *intergroup non-confronting group*) did not confront the remark (n=75 continued without reply, n=37 had non-confronting responses), and the rest confronted (24%, n=37), agreed (n=4), were suspicious (n=2), or unidentifiable (n=1). In the interpersonal condition (n=159), 89% (n=142, *interpersonal non-confronting group*) did not confront the remark (n=65 didn't reply, n=77 non-confronting responses), and the rest confronted (11%, n=17).

We found no baseline differences between intergroup non-confronters compared to interpersonal non-confronters or to exposure group on demographics (*p*'s>.25), political orientation (*p*'s>.20), or socio-political–intergroup orientations (*p*'s>.10). Importantly we also found no difference on confronting importance (*p*'s>.25), and feeling-thermometer towards Latinos (*p's*>.15). (Also, no differences between exposure and interpersonal non-confronting group on these measures; *p*'s>.18). However, unexpectedly, when looking at *baseline* Modern Racism Scale (MRS) there was significant difference, which was also the case when comparing *original* conditions (no exclusion). Namely, exposure and intergroup condition significantly differed (*p*=.02; Exposure vs. Interpersonal: *p*=.08). Participants in the exposure condition reported lower scores on MRS than participants in the other conditions. Due to this occurrence, we did not further focus on this scale as main outcome measure (but on feeling-thermometer),[26] and conducted main analyses both while controlling and not controlling for MRS pre-scores.

We first conducted a mixed ANOVA on *feeling-thermometer*, and found no significant interaction between time and experimental groups, see Figure 6. However, consistent with our predictions and results of studies 1-2, we found that participants in the intergroup non-confronting group expressed significantly less favorable feelings towards Latinos after compared to prior to the witnessed incident. No attitudinal change occurred among participants in the exposure condition. Nor among interpersonal non-confronters.

To test the between-group differences on the post-test measurement of feeling thermometer, trivialization and responsibility denial, we ran a three-level factor ANCOVA with Helmert contrast (in GLM), while controlling for feeling thermometer obtained at pre-test (covariate). This analysis specified two orthogonal contrasts in one model: (1) comparing intergroup non-confronting group to the average of exposure and interpersonal non-confronting groups (predicted to be

[26] Note, when analyzed, we found that MRS did not significantly change pre- to post-test among any experimental groups.

significant); (2) comparing exposure group to interpersonal non-confronting group (predicted to be non-significant).

As predicted, on *feeling-thermometer*, we found that participants felt more negatively towards Latinos in the intergroup non-confronting compared to the average score of the two other groups, (also significant when controlling for MRS pre-scores, *p*=.044; or without covariate, *p*=.006). As predicted, the exposure condition did not significantly differ from the interpersonal non-confronting group (also not significant when controlling for MRS pre-scores, or without covariate, *p's*>.25).

Regarding *trivialization*, as predicted, we found that intergroup non-confronters trivialized the prejudiced incident more than the other two groups, (also significant when controlling for MRS pre-scores: *p*=.004; or without covariate, *p*<.001). Trivialization in the exposure condition did not significantly differ from the interpersonal non-confronting group, (controlling for MRS pre-scores: *p*>.25; or without covariate, *p*=.073).

Regarding *responsibility denial*, as predicted, we found that intergroup non-confronters were more likely to deny responsibility than the other two groups (also significant when controlling for MRS pre-scores or without covariate, *p's* <.001). Responsibility denial in the exposure condition did not significantly differ from the interpersonal non-confronting group, (controlling for MRS pre-scores: *p*=.077; or without covariate, *p*>.25).

To test the boundary conditions of the proposed effect with importance of confronting as moderator. We ran a moderation analyses for a multicategorical IV (Hayes, 2018) on feeling-thermometer. The analyses involved two dummy variables as independent variables:

D2 (1 = interpersonal non-confront and 0 = intergroup non-confront and exposure; aka. intergroup non-confronting vs. interpersonal non-confronting comparison), and D2 (1 = exposure and 0 = intergroup non-confront and interpersonal non-confront; aka. intergroup non-confronting vs. exposure comparison). Variables were not z-standardized or centered. We controlled for feeling-thermometer pre-scores. We found no significant two-way interactions for the comparison of intergroup non-confronters vs. interpersonal non-confronters, and neither for intergroup non-confronters vs. exposure (*p*'s > .38). While the interaction was not significant, to test our specific prediction about those who perceive low importance to confronting, we probed the interaction with simple slopes interaction analyses. As expected, for those at low confronting importance (1 SD below the mean), there was no significant effect of condition on feeling-thermometer. As predicted, for those at the mean level of confronting importance, prejudice was higher among participants in intergroup non-confronting group. For those who scored high on confronting importance (1 SD above the mean) the effect was in the same direction but not significant. For statistical values see Appendix D.

Finally, we tested negative affect and discomfort as mediators for the effect of witnessing and not confronting prejudice (vs. control group) on feeling-thermometer and trivialization, and they were not significant. Also, the condition (intergroup non-confronting vs. interpersonal non-confronting vs. exposure) had no effect on negative affect (*p*'s>.25) nor on discomfort (*p*'s>.15).

Discussion

Findings of study 3 provided further validation to the motivated prejudice effect, although unlike in Study 1-2, the interaction between time and experimental groups on prejudice was not significant. However, according to our main prediction, repeated measures analyses indicated that only those participants *became* more prejudiced (feeling-thermometer) who witnessed and had an opportunity but did not confront prejudice. The same participants also trivialized the prejudiced event more and denied responsibility for acting, compared to the control groups. Like in Study 1, we again found reduced responsibility denial as an effect of not confronting intergroup prejudice, suggesting that while in Study 2 it was not significant, it is unlikely to be a false positive finding. We next conducted an internal meta-analyses on all measures to test the reliability of these findings.

Overall results suggest that the effect of witnessing and not confronting prejudice on intergroup attitudes cannot be explained by mere exposure to prejudice. Witnessing prejudice, without opportunity to confront (and without need for justifying prior inaction), did not alone normalize prejudice endorsement.

We found that the change in intergroup attitudes among non-confronters was (somewhat) dependent on participants' baseline perceived importance of confronting prejudice. Those lower on valuing confronting did not show the proposed effect – presumably because inaction did not contradict their personal values. Those moderate to higher on valuing confronting showed the motivated prejudice effect, because they were in need for justification for not confronting. Interestingly, for those non-confronters who were at the very high end on the importance scale the effect was also not significant. Perhaps their outgroup attitudes are held more stable and unlikely to change. Also, we perhaps cannot generalize, because some participants in this group inclined to negative attitude change while others to compensate (expressing more positive attitudes) for not confronting. Indeed, low-prejudiced individuals are often motivated to rectify and compensate prior prejudicial behavior (e.g., Dutton & Lake, 1973).

Importantly, while we observed some boundary conditions for the effect, the moderation was not significant, on which we elaborate in general discussion. Furthermore, we could not find support for dissonance with directly assessing psychological discomfort and negative affect, as the manipulation had no effect on these affective scales (we also elaborate on this in general discussion).

While we found no potential prior difference on perceived importance of confronting, feeling-thermometer, demographics or socio-political–intergroup orientations, our main outcome measure, MRS showed prior difference. However, this difference was "already there" when we randomly assigned participants to conditions. By chance, participants in the exposure condition were to begin with lower on MRS than participants in the other conditions. This is an unfortunate occurrence on which we had no influence, however it does little to affect our main results: participants who witnessed prejudice and did not confront, albeit having an opportunity, *became* more prejudiced towards Latinos, and this change did not occur in the control groups.

Table 2. Means and standard deviations of dependent variables across studies, and correlations between post-test measures (correlation only among intergroup non-confronters in brackets)

		M (SD)			**Pearson coefficient (r)**	
		intergroup non-conf	interpersonal non-conf	exposure	*trivializ.*	*respons. denial*
outgroup attitude	S1	*4.43 (1.18)* 4.11 (1.33)	*4.51 (1.16)* 4.60 (1.24)	-	-.19* (-.01)	-.28** (-.30**)
	S2	*4.10 (1.50)* 3.79 (1.45)	*4.13 (1.51)* 4.14 (1.47)	-	-.54** (-.48**)	-.36** (-.11)
	S3	*74.46 (20.51)* 70.83 (22.26)	*77.09 (21.72)* 76.45 (21.26)	*79.43 (22.74)* 78.51 (21.82)	-.23** (-.15)	-.19** (.008)
trivializ.	S1	3.89 (1.44)	3.19 (1.38)	-	1	.51** (.50**)
	S2	4.16 (1.58)	3.38 (1.63)	-	-	.57** (.62**)
	S3	3.52 (1.58)	3.07 (1.61)	2.74 (1.57)	-	.59** (.57**)
respons. denial	S1	4.61 (1.55)	4.09 (1.42)	-	-	1
	S2	4.43 (1.37)	4.01 (1.78)	-	-	-
	S3	4.67 (1.72)	3.83 (1.60)	3.98 (1.84)	-	-

Notes. Outgroup attitude measures were perceived trustworthiness of Jews in study 1 (7-point Likert scale), social closeness to Muslims in study 2 (6-point Likert scale), and feeling-thermometer towards Latinos in study 3 (0-100 continuous slider). Pre-test scores (when available) are in italics. S1 denotes study 1 and so on. For correlations, * *p*<.05, ** *p*<.01.

Table 3a. Statistical values for mixed ANOVA between experimental groups and time (pre- and post-test sessions) on outgroup attitudes, within-subjects and between-subjects effects.

		N	*F*	*p*	*etasq.*	*95% CI*
Study 1	group x time	138	**4.38**	**.038**	**.031**	
	within intergroup	75	**6.00**	**.016**	**.042**	**.06, .58**
	within interpersonal	63	0.35	.554	.003	-.37, .20
	intergroup vs. interpersonal		**5.84**	**.017**	**.041**	
Study 2	group x time	120	**4.50**	**.036**	**.037**	
	within intergroup	49	**7.44**	**.007**	**.059**	**.09, .55**
	within interpersonal	71	0.00	.971	.000	-.19, .20
	intergroup vs. interpersonal		**4.86**	**.029**	**.040**	
Study 3	group x time	410	1.54	.217	.007	
	within intergroup	112	**6.82**	**.009**	**.016**	**.90, 6.35**
	within interpersonal	142	0.27	.603	.001	-1.78, 3.06
	within exposure	156	0.61	.436	.001	-1.40, 3.23
	intergroup vs. controls			**.016**		**-6.74,-0.70**
	interpersonal vs. exposure			.871		-2.90,3.42

Note. Significant effects are bolded.

Table 3b. Between-subjects comparison between intergroup non-confronters vs. control groups (interpersonal non-confronters in Study 1 and 2, and also exposure in Study 3).

		F	*p*	*etasq.*	*95% CI*
Trivialization	Study 1	**8.30**	**.005**	**.058**	
	Study 2	**9.57**	**.002**	**.076**	
	Study 3		.001[a] .104[b]		0.22, 0.90 [a] -0.65,0.06 [b]
Responsibility denial	Study 1	**4.14**	**.044**	**.030**	
	Study 2	2.17	.144	.018	
	Study 3		<.001 [a] .335 [b]		0.33, 1.07 [a] -0.20,0.58 [b]

Note. Significant effects are bolded. [a] comparison between intergroup non-confronters and the average of control groups. [b] comparison between exposure and interpersonal non-confronters.

Figure 6. Outgroup attitudes as a function of experimental groups and pre- and post-test sessions (time) across studies. (Error bars *95% CI*.)

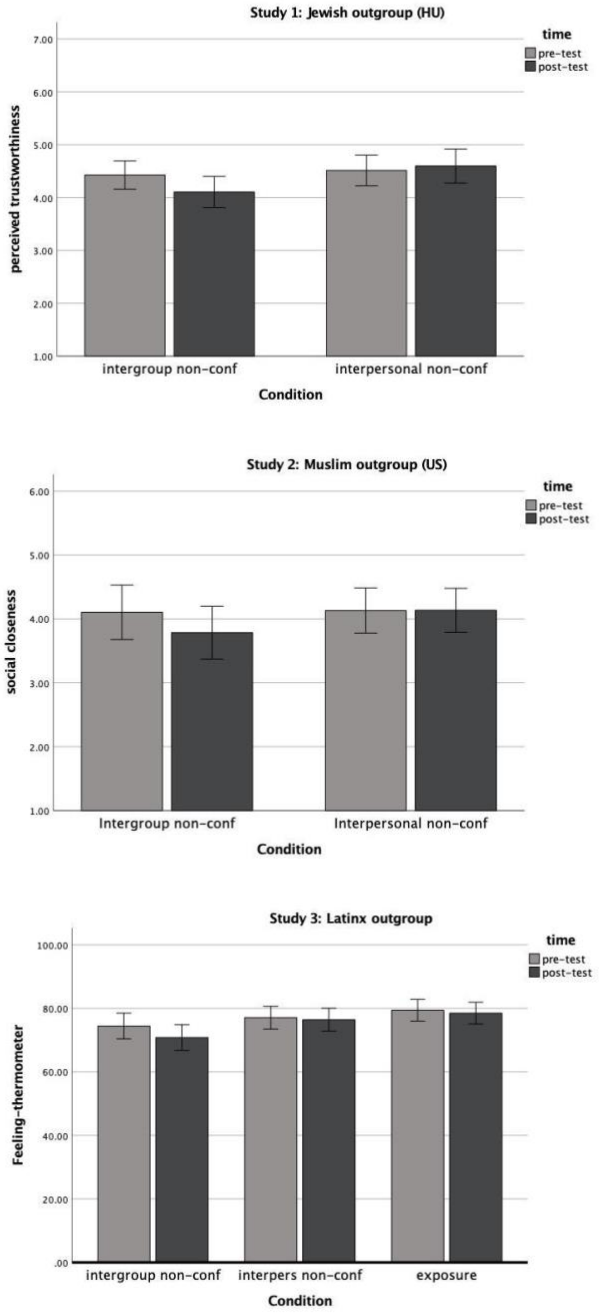
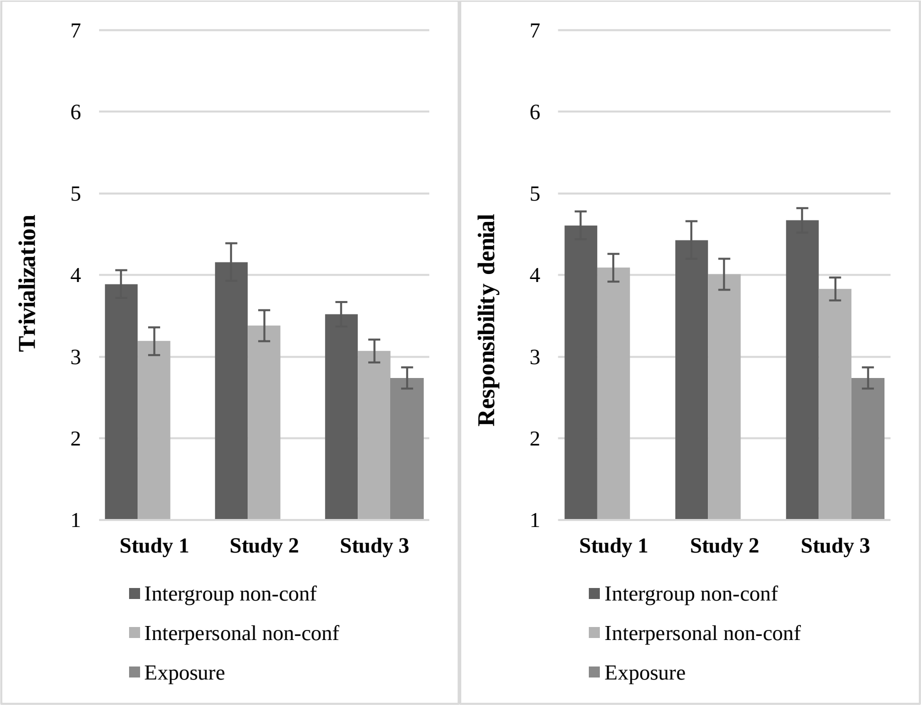


Figure 7. Trivialization (of the intergroup prejudiced event) and Responsibility denial (for acting in the situation) as a function of experimental groups across studies. (Stander Error bars).



Internal Meta-Analysis

To examine the robustness or consistency of our findings, we conducted internal meta-analyses (following Goh et al., 2016) on outgroup attitudes (perceived trustworthiness in Study 1, social distancing in Study 2, and feeling-thermometer in Study 3), on trivialization, and on responsibility denial. We meta-analyzed all three studies (N = 668) using fixed effects in which each key effect size was weighted by sample size. We found that the interaction between time (pre vs. post) X condition (intergroup vs. interpersonal in study 1-2, and vs. exposure in study 3) on outgroup attitudes was significant (M*r* = 0.12, *Z* = 3.15, *p* < .005, two-tailed). The within-subjects overtime change among intergroup non-confronting group on outgroup attitudes was significant (M*r* = 0.18, *Z* = 2.67, *p* < .01, two-tailed), suggesting it is a robust effect. The effect of intergroup non-confronting vs. control conditions on outgroup attitudes, while controlling for pre-scores, was also significant (Mr = 0.13, Z = 3.22, p < .005, two-tailed), suggesting that this effect is robust. Similarly, for trivialization the effect was also significant (Mr = 0.19, Z = 5.03, p < .0001, two-tailed), suggesting this is also a robust effect. On responsibility denial, the effect was also significant (Mr = 0.17, Z = 4.37, p < .0001, two-tailed).

General Discussion

In the current research, we demonstrated a path through which prejudice perpetuates and intensifies over time and found evidence for what we termed, a *motivated prejudice effect*. We tested how witnessing, and not confronting prejudice and discrimination (although having an opportunity to) changes the non-confronters' intergroup attitudes for the worse. We conducted two pilot studies and three experiments using online simulations we developed for participants to engage in. We tested our predictions across multiple intergroup contexts where the outgroup minority in the US was either African American, Muslim American or Latinx American, or Jewish in Hungary. Additionally, across our main studies, we used a mixed within- and between-subjects design, where we assessed participants both prior and following witnessing the biased event (pre- and post-test). This design enabled us to test overtime changes in prejudice among those who did not confront, and to compare those changes to control groups.

We predicted and found that those participants who witnessed intergroup prejudice and had an opportunity to confront the perpetrator, but did not do so, endorsed more negative intergroup attitudes relative to their own attitudes prior to the incident (Studies 1–3). Additionally, they showed more negative attitudes than those who did not witness any bias (Pilot studies), or those who witnessed and did not confront other (non-intergroup) type of bias (Studies 1–3). Importantly, there was also no overtime attitude change for this latter control group (Studies 1–3). Comparison to interpersonal group helped to rule out the possibility that the effect of not confronting prejudice is simply a derogatory response resulting from deflated self-esteem that would be triggered by any personal failure (Fein & Spencer, 1997) of not confronting mistreatment of others. It also rules out the explanation that the observed change in attitudes would be specific to a personality type, who does not confront in general.

Furthermore, we found that those who did not confront prejudice had worse intergroup attitudes than those who witnessed the same intergroup prejudice scenario but did not have a chance to confront (exposure condition), therefore likely felt no need to justify their inaction (pilot study 2 and Study 3). Importantly, the exposure control group also did not show overtime change in their outgroup attitudes (Study 3). This condition enabled us to isolate the effect of opportunity to confront, which is central for the attitude change to occur and provides further support for the dissonance-reduction account. This way we ruled out the possibility that the observed change among non-confronters was simply due to being exposed to prejudice. That is, contrary to what previous work would suggest, the effect was likely not triggered by desensitization, or persuasion or change in the normative context (e.g., Blanchard et al., 1991; 1994), and it was not a product of just-world beliefs to blame victims (e.g., Janoff-Bulman & Frieze, 1983; Lerner & Simmons, 1966; Lerner & Miller, 1978).

Instead, we suggest that our findings can be explained by need for cognitive consistency, such as through a dissonance-induced self-justification mechanism (Abelson et al., 1968; Festinger, 1957; Rasinski et al., 2013), whereby people felt an inconsistency between their cognition (prejudice is wrong and should be contested) and their inaction in face of prejudice, and they were motivated to reduce this dissonance. Given that their past behavior cannot be changed (i.e., their inaction in face of prejudice), they resorted to changing the relevant cognition that was at odds with their behavior. In accordance, we found that participants changed their attitudes about the outgroup (i.e., come to rationalize that there is some truth in the expressed sentiment), about the incident (i.e., believe that it was not that harmful), and about their responsibility in the situation. Namely, we observed (prejudicial) attitude change (studies 1-3), trivialization (studies 1-3) and responsibility denial (study 1 and 3), which are actually prevalent dissonance-reduction strategies (McGrath, 2017).

Beyond prejudice, trivialization and responsibility denial is harmful to intergroup relations as it can build tolerance for prejudicial atrocities in the long-run which in turn likely to go uncontested. Note that we had no a-priori prediction in reconciling the relationship between these different modes of dissonance-reduction. Looking at the correlations among intergroup non-confronters, we saw inconsistent, moderate to no association between prejudice, trivialization and responsibility denial across our studies (see Table). As suggested in the literature, it is possible that people engage in more than one strategy for the same dissonant event (e.g., Elliot & Devine, 1994), but researchers still know little about simultaneous and autonomous use of strategies (for review see McGrath, 2017). Future research should investigate the relationship between different modes of dissonance reduction in such context.

In order to further investigate the dissonance-justification account, we tested whether the effect of witnessing and not confronting prejudice on intergroup attitudes would be dependent on participants' baseline perceived importance of confronting prejudice (Study 3). Results on boundary conditions of the proposed effect partially support this account. Namely, those who did not value confronting did not show the motivated prejudice effect, likely because inaction did not contradict their personal values, thus they did not seek justification for not confronting. While we observed such boundary conditions for the effect, the moderation (interaction) effect was not significant. This could be because most participants felt some need to justify inaction given the explicitly unfair (non-meritocratic) treatment or given their sole responsibility to act (the prejudiced player addressed them directly with his remark) or given that they had little external justification for not acting (an online interaction with minor financial consequences). Additionally, the perceived importance of confronting scale was explicit, and it might have triggered socially desirable responses, thereby providing a biased estimate. Indeed, the median for confronting importance scale in the sample was relatively high, 7.2 on a 10-point scale.

While inferring dissonance reduction process through its corresponding attitude change is in accordance with conventional practices of indicating this (mostly theorized) construct (e.g., Cameron & Payne, 2012; Leippe & Eisenstadt, 1994; Rasinski et al., 2013), future research employing physiological measures could best investigate this potential explanation (although some argue that physiological arousal is a questionable proxy for affective arousal, e.g., Satpute et al., 2019). We aimed to directly show the effect of dissonance-induced discomfort by using self-report affective scales (psychological discomfort and negative affect; Study 3). However, we found no significant differences between experimental conditions on these measures. We assume that we failed to find any effect for one, because it is challenging to capture such delicate discomfort sensations with self-reports (previous work on dissonance also did not find effect on discomfort scale; Gosling et al., 2006). Additionally, because the time at which non-confronters responded to these measures may already had been subsequent to the intrapersonal dissonance-reduction process, thus responses on these scales could no longer detect this discomfort. We believe that lack of such evidence does not conclude that dissonance-reduction does not account for our findings, but rather that methodology is limited in identifying it.

One could argue that attitude polarization may explain our findings, which would suggest that a person’s initial attitudes and opinions can strengthen and intensify after exposure to either supporting or opposing views (especially if one needs to defend those views, e.g., Chaiken & Yates, 1985; Myers, 1975; Myers & Lamm, 1976, Saguy & Szekeres, 2018). However, for one, we found no initial differences between the non-confronters and control groups on socio-political or prejudicial attitudes. Findings among those participants who confronted intergroup bias also rule out this explanation. Specifically, exploratory analyses revealed no significant overtime change among intergroup confronters (study 1 and 3, where available; *p*’s > .25). Moreover, those who perceived low importance in confronting prejudice did not show the proposed effect, that is, those likely most prejudiced did not become more prejudiced following exposure to prejudice. These rule out attitude polarization as a likely explanation.

Self-perception theory (Bem, 1967) could also explain our findings and suggest that non-confronters inferred their intergroup attitudes from the observation of their own (non-confronting) behavior when facing prejudice. However, self-perception process applies to situations where internal cues are weak and ambiguous, so individuals need to rely on external cues in order to understand their own attitudes (Bem, 1967). While we did not measure attitude strength per se, the fact that the motivated prejudice effect did not occur exactly among those who scored low on perceived importance suggest that self-perception is not the most likely explanation. Overall, whether self-perception or dissonance is the correct explanation for such attitude changes is anyway a source of great controversy in the literature, many suggesting that it is nearly impossible to distinguish (e.g., Greenwald, 1975), and we similarly are unable to determine with certainty. *Overall*, while we might not yet have direct evidence for the underlying mechanism, we found increased negative outgroup attitudes which followed (only and solely) from witnessing and not confronting prejudice when having an opportunity to – which in itself is a significant phenomenon that warrants attention.

Indeed, considering our findings, one is left wondering why participants did not compensate instead, not even those who highly valued confronting prejudice. Specifically, why participants tended to endorse more negative outgroup attitudes as means to resolve their discomfort instead of tending to rectify their failure and compensate the outgroup (e.g., especially in the Pilot studies participants could have offered more money to the Black organizations – instead of less, as they ended up doing). Certainly, people sometimes act morally to compensate for prior immorality (Sachdeva et al., 2009). For example, when participants' low-prejudiced identity was threatened (with false feedback) they were later more generous to a black panhandler than participants whose identity was not threatened (Dutton & Lake, 1973).

On the other hand, our findings are not surprising given prior research suggesting that ingroup members are often more likely to derogate or dehumanize an outgroup instead of compensating them, in reaction to the ingroup's wrongdoing against that group (Castano & Giner-Sorolla, 2006; Glasford et al., 2009). When light is shed on people's own moral failures they similarly tend to act defensively, which prevents them to rectify their behavior (e.g., Gausel & Leach, 2011). Approaching the question practically, perhaps participants in our study did not take the opportunity to compensate because they felt a potential donation would not sufficiently make up for not confronting (as literature on dissonance reduction strategies selection would suggest; McGarthy, 2017). While answering this question was beyond the scope of the present study, it is a fruitful avenue for future research.

In the present research, we employed an online paradigm, rather than in-lab studies, due to ethical considerations (an online setting minimizes the emotional obtrusiveness of such disturbing situations) and to ease the feasibility of conducting multiple studies in different intergroup contexts with required sample sizes. The external validity and generalizability of the findings is limited in this respect. There are some differences given an in-person experience, such that the situation may feel more shocking, responsibility to act more emphasized, less external justification for not acting (offline confronting may be considered more effective), or actually more external justification (e.g., avoiding physical attack). Yet we assume that the motivated prejudice effect is not specific to online context and would also occur in-person (more about this in Chapter 4). Nevertheless, in future research, the proposed effect should be tested in such context. Additionally, conducting the study online both strengthened internal validity of the findings (computerized survey, all participants were exposed to the exact same stimuli), however it also weakened it to the extent we were not able to control their physical surrounding when participating in the study. However, we can assume that such individual differences were spread across conditions, and so random assignment reduces this weakness.

Additionally, there is the question of differing confronting rate, as there was an apparent difference in confronting level between the Hungarian (8%) and US sample (24–42%). On the one hand, a possible explanation is the hurtfulness of the prejudicial slur which was very blatant and hostile in the U.S. studies, and in Hungary the comment was subtle. The differing confronting rate may reflect cultural differences, for example in assertiveness, although in Hungary confronting was relatively higher, 22% in the interpersonal condition. Another explanation may be differences in norms regarding prejudice and prejudice confrontation. However, we have some indication of this norm across the intergroup contexts. Mean (and median scores) on trivialization of the event among control groups were below midpoint in scale (and at similar levels) across studies. Also, we asked control groups if they would have confronted the (prejudiced) player by messaging him, and *hypothetical* confronting rates were similarly (around or) above midpoint in scale across studies (see Appendix C). Indicating that participants on average, in both countries, believed that this behavior was serious, not acceptable and warrants confronting. Nevertheless, it is possible that there are cultural differences in how people act (how brave they are) when placed in the actual situation. Additionally, there may be difference in the *perceived* societal norm of prejudice, that is, to what extent participants believe that the expressed prejudice is acceptable in their society, and whether the majority of people would confront or not in this situation. Future research should explore how this perceived norm would influence confronting rate. In general, we argue that the replication of the motivated prejudice effect in two quite different cultures (especially in regard to prejudicial norms), and across multiple different intergroup contexts, is a strength of our research as it contributes to the generalizability of our findings.

The witnessed incident and measures were framed around intergroup trust and liking (in studies 1-3), however the actual slurs varied to fit the predominant prejudice about the target outgroup, and outgroup attitude measures varied accordingly (pilot studies and study 1) and also by keeping in mind what is used in the literature (mostly considered for study 2-3). This led to using different scales in each study, which on the one hand increases external validity given that the effect was found across different measures, yet it is also problematic in terms of methodological scrutiny and replicability.

Finally, we ought to acknowledge that there is a selection issue because we exclude those who confront from an experimental condition. This is an inherent methodology problem in our research question about those who do not confront, and it weakens the internal validity of the studies. Besides our attempts to alter the online paradigm to decrease confronting (similarly to Rasinski et al., 2013), the pre-posttest design is the most optimal approach to address this issue. We had to give participants an opportunity to confront, because if we provide them with sufficient external justification for not confronting, they will not experience dissonance. Like in real-life, it is people's own choice to confront or not and this decision may have consequences for those (specific) people's prejudicial attitudes, and we aimed to test these particular consequences.

Conclusion

In the present research we shed light on intergroup costs for staying silent in face of prejudice and discrimination. We showed that when non-target bystanders do not confront, not only they miss an opportunity to challenge the views of the perpetrator, but they themselves subsequently come to endorse the expressed prejudiced sentiment. We identified a route via which prejudice (not confronted) not only perpetuates but exponentially amplifies in a given social environment. Given the growth of diverse societies and the occasional simultaneous rise in prejudiced discourse and atrocities (Craig & Richeson, 2014), potential bystanders in the context of prejudice are becoming increasingly common, rendering the focus of the present research timely and relevant.

Chapter 3: When it’s your loss – The effect of moral loss and gain mindset on confronting prejudice

This chapter is based on:

Szekeres, H., Halperin, E., Kende, A., & Saguy, T. (2019). The effect of moral loss and gain mindset on confronting racism. *Journal of Experimental Social Psychology*, *84*, 103833.

Introduction

Elin Ersson, a young university student boarded a plane heading from Sweden to Turkey in July 2018 to protest the deportation of an Afghan asylum seeker, who was forced on that flight. She refused to sit down, preventing take off, until the man was removed from the aircraft. Ersson sacrificed a lot. During the protest she faced an angry cabin crew, complaints by other passengers, and potential legal charges. Eventually, Ersson succeeded in her protest and received an ovation from passengers and was widely praised on social media around the world for her intervention, many calling her a hero. The current research addresses the motives and the mechanisms underlying such confronting behavior by third-party individuals. We investigated how a prospective personal moral failure (of not intervening), or moral gain (of intervening) can play a key role in motivating people like Ersson to perform courageous acts.

Moral courage is a willingness to take a stand in defense of one’s own moral principles even when others do not (Miller, 2000; Skitka, 2012). Ersson’s action is unique to the extent that the majority of people think they would speak up against intergroup bias, however, often they do not. For example, even though White Americans anticipated feeling upset and taking action against someone who espouses racial bias against a Black person, those put in that actual situation reported little negative affect and forewent punishing the racist person (Karmali et al., 2017; Kawakami et al., 2009). Similarly, heterosexual participants who imagined witnessing a homophobic slur reported higher intentions of confronting than people who actually witnessed the slur (Crosby & Wilson, 2015). Such inaction is unfortunate because confronting, especially if done by third-party individuals, can actually change perpetrators’ beliefs and reduce prejudice (Czopp & Monteith, 2003; Czopp et al., 2006).

Research suggest that people fail to exhibit moral courage and confront intergroup bias because they rather avoid interpersonal costs, such as being perceived as a troublemaker, being disliked or experiencing retaliation (e.g., Kaiser & Miller, 2001, 2003; Eliezer & Major, 2012;

Shelton & Stewart, 2004; Swim & Hyers, 1999). While it has been seldom discussed in the confronting literature, omission of intervening can entail personal, psychological costs as well, in the form of deflating one's moral self-concept. Indeed, maintaining a non-prejudiced self-image is important to many individuals (Dutton & Lennox, 1974; Monteith, 1993; Plant & Butz, 2006; Plant & Devine, 1998). When individuals' non-prejudiced identity is threatened, they employ different strategies to reinstate it, such as engaging in downward social comparisons with bigots (O'Brien et al., 2010; Wills, 1981), inhibiting prejudiced responses to jokes (Monteith, 1993), or being more generous to an outgroup member (Dutton & Lake, 1973). Thus, we can expect individuals who identify as non-prejudiced to be motivated to confront racism when feeling their moral identity is at risk - which likely is the case when they witness racism and have an opportunity to intervene (i.e., confront). Along the same lines, confronting can also provide personal gains by enhancing one's moral self-concept. Accordingly, individuals might be motivated to confront racism because they construe it as an opportunity for self-enhancement (i.e., desire to increase positive self-concept; Leary, 2007) or self-improvement (i.e., desire to improve aspects of one's self-concept; Sedikides, 1999; Sedikides & Hepper, 2009).

Thus, we here propose that considerations about one's own morality likely weigh in when deciding whether to confront prejudice or not. Such considerations can be induced in different ways, even by simply priming aspects of moral courage. For example, people associate different moral behaviors with different moral prototypes (helping with being caring, moral courage with being just, heroism with being brave). Accordingly, an activation of a certain prototype (e.g., "just", which is associated with being fair, moral, truthful, honest) was shown to increase the tendency for morally courageous behavior (Osswald et al., 2010). Extending previous research, we go beyond the moral priming effect to investigate the unique role that moral loss and moral gain mindsets potentially plays in motivating moral behavior, like confronting. According to our thinking, a person with a loss mindset is likely to feel that by not confronting bias, he/she can lose a sense of moral integrity. A person with a gain mindset is likely to feel that by confronting, he/she will earn a sense of being a more moral human being.

While both a moral loss and moral gain mindsets are related to one's moral self-concept, they are also psychologically different and accordingly may motivate confronting in different ways and for different people. Corresponding to these two mindsets, self-regulatory focus theory (Crowe & Higgins, 1997; Higgins, 1997) distinguishes between two motivational systems that regulate goal-directed behavior: a *promotion* and a *prevention focus.* Promotion focus emphasizes advancement and growth, with goals being viewed as ideals (Shah & Higgins, 1997). Prevention focus emphasizes safety, duties and responsibilities, with goals being viewed as obligations. Those with promotion focus are primarily concerned with the presence or absence of positive outcomes (or end states), while those with prevention focus are concerned with negative outcomes. Thus, promotion focus orients people on pursuing opportunities, whereas prevention focus orients toward avoiding errors (Crowe & Higgins, 1997; Higgins et al., 1994).

Messages that promote a moral gain mindset regarding confronting (e.g., "intervening would reveal a good and moral side of me") correspond to a promotion focus to the extent that it emphasizes the opportunity of a positive moral end state if the person confronts racism. Similarly, a moral loss mindset (e.g., "not intervening would reveal a bad and immoral side of me") corresponds to prevention focus to the extent it orients people to avoid making an error and ending up with a negative moral state if one fails to confront racism.

Research on regulatory focus suggests that under a prevention focus people are likely to react more strongly to issues related to justice and morality (Sassenberg & Hansen, 2007), especially if they are morally committed to that particular goal. For example, when individuals were primed with a prevention focus (wrote about what they felt they ought to achieve in their working life), the more they held a moral conviction about the fair treatment of their group, the more they supported collective action against ingroup discrimination (Zaal et al., 2011). Researchers argued that because prevention-orientation makes people construe goals such as those mandated by moral conviction (in this case, fair treatment) as necessities (Scholer et al., 2010; Shah & Higgins, 1997; Zaal et al., 2012), they were, presumably, particularly sensitive to the possible losses of inaction and were motivated to avoid those. Meanwhile, for individuals primed with a promotion focus (who wrote about what they would ideally like to achieve) moral conviction did not predict collective action intentions (Zaal et al., 2011). The researchers assumed that promotion-oriented individuals, for whom expectations for success play a key role in taking action (Shah & Higgins, 1997; Zaal et al., 2012) were likely doubtful regarding the effectiveness of collective action. Thus, this work exemplifies how a loss mindset and a gain mindset trigger a different set of concerns, resulting in different actions.

Another study exploring the relationship between moral commitment, regulatory focus and moral behavior found similar results (Brebels et al., 2011). Business students were made to imagine being managers of a company, and their procedural justice intentions were assessed. Results revealed that participants for whom morality was a central part of their identity, exhibited more procedural justice intentions under a prevention focus (manipulated via priming a threat to the company's position in the market) than under a promotion focus (manipulated via priming an opportunity to advance the company's position; Brebels et al., 2011). One possible explanation is that under a prevention focus, participants focused on the possibility of feeling immoral (loss to moral identity), which they were motivated to avoid. Those for whom moral identity was less important, showed the opposite pattern, i.e., more justice intention in promotion than in prevention focus (Brebels et al., 2011). Perhaps under a promotion focus they were made to think about how they could potentially improve their moral identity by acting fairer.

Together, the work described above suggest that a loss mindset is likely to promote intentions for moral behavior, particularly for those who care about being moral. Applied to our context, this suggests that a loss mindset can promote confronting among those highly committed to non-prejudice, while potential gains might not. This notion echoes prospect theory (Tversky & Kahneman, 1991; 1992), according to which losses inflict psychological harm to a greater degree than gains gratify, which means that people are more willing to run risks to avoid losses than to approach gains. Thus, the psychological costs of falling short of one's moral self-concept should be a motivating force in confronting racism for those who care about being non-prejudiced.

Nevertheless, a loss mindset is not likely to cause change in confronting rate among those weakly committed to non-prejudice, because they should perceive little threat to their non-prejudiced (moral) self-concept as a result of not contesting racism. On the contrary, they might even appraise a loss message as external pressure and obligation to respond without prejudice and thus reduce their intention to confront as a result of a backfire effect (Does et al., 2012; Legault et al., 2011; Powell et al., 2005).

Meanwhile, a focus on gains to one's moral self-concept could drive more confronting among those weakly committed to non-prejudice because it is seen as an opportunity to improve moral self-regard (Leary, 2007; Sedikides, 1999; Sedikides & Hepper, 2009). Such opportunity for moral self-improvement should play less of a role in motivating confronting among those who already view themselves as non-prejudiced. Additionally, under moral gain focus, those weakly committed to non-prejudice may confront to gain moral credits prospectively in the domain of racism (Cascio & Plant, 2015) – indeed, prejudiced individuals show higher tendency than non-prejudiced individuals to license their biased/immoral behavior with prior unbiased/moral behavior (Effron et al., 2009).

The present research

Taken together, we predicted that participants' moral commitment to non-prejudice would moderate the effects of moral mindset on confronting racism. Specifically, a moral loss (vs. control) mindset would significantly increase confronting tendencies among those strongly morally committed to non-prejudice, but not among those weakly committed (H1). We also predicted that a moral gain (vs. control) mindset would drive confronting among those who are weakly committed to non-prejudice and would not affect those strongly committed (H2).

We tested these hypotheses in two experiments. In the first, participants were asked to picture themselves in specific racist scenarios, for which confronting was framed in one of three ways: as moral loss, as moral gain, or neither. Participants were asked to report their willingness to confront. In the second study, participants were again randomly assigned to either a loss mindset, a gain mindset, or an empty control using a different manipulation that we considered to be possibly more enduring. After a few days they went through a behavioral paradigm where they witnessed racism and had an opportunity to confront. By investigating whether moral mindsets increase confronting, the present research allowed us to gain insight into different motivations for confronting racism, which can in turn inform interventions aimed at promoting standing up against racism, as well as against other forms of immoral behavior.

Based on the literature on regulatory focus and moral orientation (Brebens et al., 2011; Zaal et al., 2011) we assessed participants' moral commitment to non-prejudice with the moral identity self-importance scale (Aquino & Reed, 2002), which comprises two dimensions/subscales (internalized vs. symbolized), and with the moral conviction scale (Skitka & Morgan, 2014). Both were slightly altered to the prejudice domain. For an individual with a strong moral identity, moral strivings are integrated with the self-concept and are central to a person's self-definition (Aquino & Reed, 2002). The internalized component captures a personal and private aspect of moral self-concept, and the symbolized component captures the social and public aspect. Higher scores on moral conviction capture individuals' strong and absolute belief that something is wrong or right (Skitka, 2010).[27] We are unaware of previous work that tested these

[27] In addition, we included a measure that intended to capture the strength of one's *general* moral ought vs. ideal orientation (which reflects the tendency to avoid moral failure vs.

scales together, especially not in the current context. We treated each of the scales (3 in total) as a potential moderator of the predicted effects, and our research was exploratory as to under what condition and which morality orientation would influence the relationship between loss/gain mindsets and confronting.

Study 4: Vignette study

To provide initial support for our predictions, in this study we manipulated moral mindset and measured the self-reported tendency to confront racism. All participants were provided with two vignettes, each depicting an instance of racism (one against a Spanish-speaking boy and another against a Muslim woman, both placed in the US). They were asked to imagine themselves as taking part in these situations, namely, witnessing racism and having opportunity to confront. The opportunity to confront (or not) was manipulated to involve potential moral self-concept loss, moral self-concept gain or neither. In this latter control group participants were exposed to the same scenarios, and to the opportunity to confront, but no manipulation of loss or gain was added. This enabled us to test the effect of loss and gain beyond a morality priming effect. Then we asked participants about their willingness to confront in the situations described. As hypothesized moderating variables, participants responded to scales measuring the strength of their moral commitment to non-prejudice.

Method

Participants and procedure. We recruited 480 U.S. residents (North-American) participants via Amazon's Mechanical Turk (mTurk), who participated in an online study for monetary compensation ($1.20). Power analysis revealed that to detect an (assumed) small effect size (0.2) to achieve a power of at least 0.80 for a moderation analysis (to calculate,

approach moral ideals, respectively; Aoki, 2015) in order to demonstrate that the predicted mechanisms are specific to moral stance about prejudice and not to general morality. Given that the outcome measure was specific to intergroup situations involving confronting prejudice, we did not expect general moral orientation to moderate the effects. Results supported this expectation, see details in Appendix G. In both studies, we also included an SDO and IMS-EMS scale for exploratory purposes (see Appendix F).

we used F test ANOVA/interaction with 6 groups), the suggested sample size was 400 (G*Power 3.1; see Faul et al., 2009). Given that some participants might be excluded based on attention checks, our collected sample size was 480. Participants were randomly assigned either to the moral loss, moral gain or control group. Participants filled out the morality scales and completed the vignette scenario measures. The materials were counterbalanced such that the morality scales appeared either prior to, or after, the manipulation. At the end of the study, participants responded to demographic questions (*age*, *gender*, *education level*, *conservative–liberal orientation*; *relative socio-economic situation*, See Appendix E for full demographics), were debriefed and thanked for their participation.

Data of participants who failed the attention check (which was placed close to the morality scales, "For this question mark number seven as a response"; n = 33) were excluded, leaving 447 participants for analyses (n = 147 in loss, n = 143 in gain, n = 157 in control condition; 47.3% female, 52.2% male, 0.5% other, M_{age} = 36.38 years, SD_{age} = 11.56, range: 19-82).[28]

Stimuli and measures. For all measures, participants responded on a 9-point scale ranging from 1 (not at all true of me) to 9 (completely true of me) unless indicated otherwise.

Moral conviction. We used the 4-item moral conviction scale (Skitka & Morgan, 2014). To reflect conviction about prejudice we gave participants the following

instruction stem: "To what extent is your position on standing up against prejudice and discrimination...". Participants then responded to the following four items: (1) "a reflection of your core moral beliefs and convictions?" (2) "connected to your beliefs about fundamental right and wrong?" (3) "based on moral principle?", and (4) "a moral stance?" Mean scores on these items were calculated for each participant composing an

[28] *Socio-economic status* (1=destitute to 6=wealthy): *M*=3.39, *SD*=1.01, destitute (1.9%), poor (16.6%), so-so (37.1%), good (31.5%), better than most (11.2%), wealthy (1.9%). *Education*: less than high school (0.7%), high school diploma (34.5%), Bachelor's degree (47.6%), Master's degree (12.1%), PhD (1.6%), Other (3.5%).

internally consistent scale (α = .94). Higher scores indicated higher levels of moral conviction against prejudice.

Moral–prejudice identity. We used the moral identity (MID) self-importance scale (Aquino & Reed, 2002) and adapted it to the current context by revising the scale instruction to refer to non-prejudiced values and by excluding one item that did not fit the current context. For the MID internalization subscale participants responded to 5 items, such as “Being someone who has these views and beliefs [being non-prejudiced] is an important part of who I am” (α = .83). For the MID symbolization subscale participants responded to 4 items, such as “The fact that I have these views and beliefs [being non-prejudiced] is communicated to others by my membership in certain organizations” (α = .92, see Appendix E for full scales).

Vignette scenarios and confronting intentions. The scenarios and measures were developed for the purpose of the current study. All participants were presented with the same two scenarios (the order of the two scenarios was counterbalanced). In scenario A, the participant had to imagine that she or he witnesses a man verbally assaulting a Spanish-speaking teenage boy on the bus and expressing his dislike of immigrants (see Appendix E for full texts). In scenario B, the participant allegedly overheard his or her co-workers making fun of their Muslim female co-worker for her religion.

For each scenario, we mentioned a dilemma and asked participants to imagine the pro’s (e.g., “you believe that this specific boy is treated unfairly, you are debating whether to intervene or not”) and con’s of confronting (e.g., “if you get involved, the man may verbally or even physically attack you”). Until this point all participants read the same scenario. The manipulation of moral loss vs. gain was communicated at the end of this text. Specifically, in the loss condition, additional arguments referring to moral considerations framed around losses were presented (e.g., “You feel it is your moral obligation to intervene. If you don’t intervene, you fail your moral duty, and you may later feel like a worse person morally. You feel you can lose a lot if you don’t confront.”).

In the gain condition the additional arguments referred to a moral gain (e.g., “You feel it is your moral aspiration to intervene. If you intervene, you succeed to live up to your moral principles, and you may later feel like a better person morally. You feel you can gain a lot if you confront.”). In order to encourage participants to carefully read the scenarios we included an open-ended question under each scenario, which read: “Based on the text, what are the considerations in the decision to intervene?”. Responses to this question were not analyzed.

Following each scenario, all participants were asked about their confronting intentions: First, we measured willingness to engage in specific confronting actions with six items rated on a 9-point scale (from 1 = not likely at all to 9 = very much likely), such as “I would confront the man and tell him he is racist.” or “I would ask the man to stop assaulting the boy.” (scenario A) and “I would tell my supervisor about my co-workers’ conversation” or “I would ask my co-workers to stop insulting her” (scenario B; see Appendix E for full measure).[29] Then, for each scenario, we also included an item assessing overall confronting willingness: “Overall, to what extent you would confront in this situation in order to [help the boy/stand up for her?]” on a 9-point scale from 1 = I would not confront at all to 9 = I would totally confront. Given that people may vary in the form of confronting they choose to take, and a participant may prefer one way of confronting very much while not at all another, we extracted the highest score each participant gave across the 6 items for each scenario (reverse coded the non-confronting option items). This way, we captured the greatest tendency to confront, for each participant. The two values (maximum value from each scenario) were then averaged with the overall general confronting scores (2 for each scenario) given by each

[29] As the seventh item on these blocks we put: “other suggestion (not mandatory, if you don't write, just mark 1): _________(text entry)”. To avoid anchoring, we did not write an item, which would describe an action in agreement with the racist perpetrator (i.e., insulting the boy or the female co-worker). We included this item to provide an opportunity for participants to express this sentiment if they wished. We did not analyze these responses or considered scores on this item in data analyses.

participant. These four numbers formed an internally consistent ‘confronting intentions’ measure (α = .75).

Following the vignette scenarios and intention measure, we asked participants in the loss and gain experimental conditions to indicate “Which of the following is closer to what was suggested in the texts about feelings and morality?” Answer options were either 1 = “After confronting, people may feel better and gain positive moral identity” (indicating moral gain) or 2 = “After not confronting, people may feel worse and loose positive moral identity” (indicating moral loss). We considered this a manipulation check.

Results

Preliminary analyses. Using One-way ANOVA and post-hoc tests we did not find significant differences between conditions on demographic variables, *p*’s > .14 (see Table 4 for means, standard deviations and correlations between study variables). However, there were significant differences between conditions on the moral conviction scale (loss vs. control: *p* = .019, gain vs. control: *p* = .004), *F*(2,444) = 4.74, *p* = 009. There were no significant differences on the MID-symbolization scale (loss vs. gain: *p* = .075, gain vs. control: *p* > .25; *F*(2,444) = 1.66, *p* > .19), nor on the MID internalization scale (*p*’s > .10; *F*(2,444) = 1.58, *p* > .20). We also tested whether the order of study materials had an effect on confronting intentions or on the morality scales and found non-significant differences (*p*’s < .12). In addition, we tested and found no significant two-way interactions between order and condition (loss vs. gain vs. control) on confronting intentions (*p* > .25) or on the morality scales (*p* = .093 for MID-symbolization, otherwise *p*’s > .25).[30] Given these results we nonetheless decided to control for order as a covariate in our main analyses.

[30] The three-way interaction between order, condition (loss vs. gain vs. control), moral commitment to non-prejudice scales (analyzed separately) on confronting were not significant either (*p*’s > .25).

Table 4. Means (standard deviations) and correlations between study variables in Study 4.

	M (SD)	Conf.	MC	MID int.	MID symb.	Cons. -Lib.	SES	Edu.
Confronting intentions	*7.02 (1.63)*	-						
Moral conviction	*7.29 (1.76)*	.43**	-					
MID internal	*7.33 (1.67)*	.26**	.53**	-				
MID symbol	*5.28 (2.23)*	.29**	.37**	.31**	-			
Conservative –Liberal	*6.29 (2.96)*	.18**	.21**	.25**	.06	-		
SES	*3.40 (1.01)*	.11**	.04	-.13**	.18**	-.14**	-	
Education	*2.79 (0.73)*	.11*	–.01	-.05	.12*	.05	.30**	-
Age	*36.38 (11.56)*	-.03	.10*	.19**	.01	-.10*	-.16**	.07

Note. * p < .05, ** p < .01. Confronting intentions, moral conviction (MC), and MID's were on a 9-point continuous scale. Conservative-liberal dimension was on a continuous slider from 0 (conservative) to 10 (liberal). SES was on 6-point and education was on a 5-point ordinal scales.

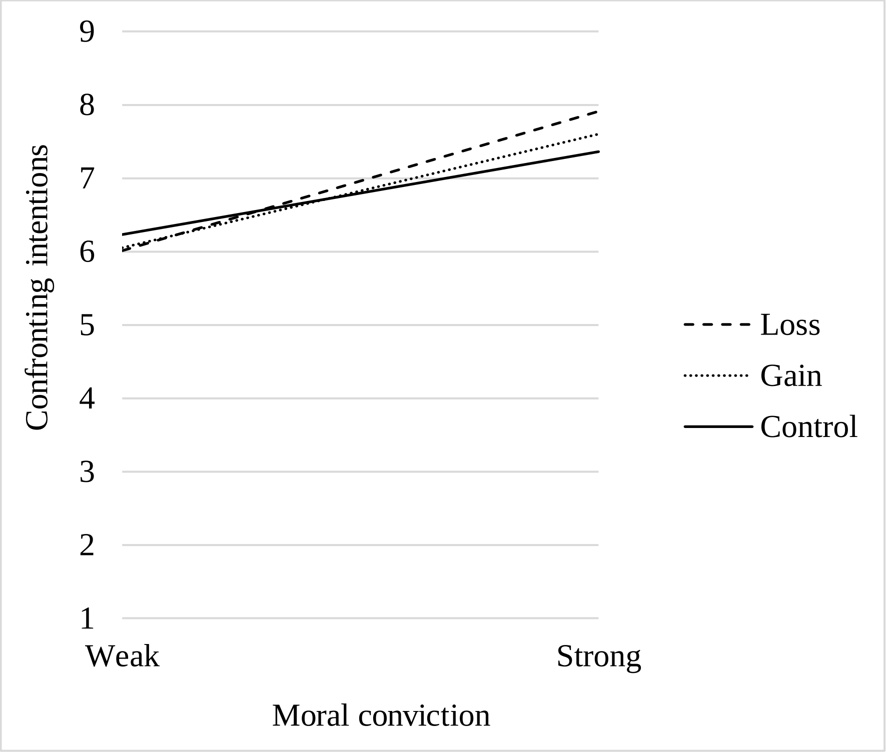
As a next step, we tested the manipulation check item. As intended, we found that in the loss condition, significantly more participants chose the loss response ("… people may feel worse and loose positive moral identity"; 66.2%) over the gain response, and in the gain condition, significantly more participants indicated the gain response over the loss response ("…people may feel better and gain positive moral identity"; 91.5%); $\chi^2(1) = 98.96$, $p < .001$, Cramer's $V = .60$.

Main analyses. In order to analyze the effects of a loss mindset (vs. control) and of a gain mindset (vs. control) on confronting intentions as a function of participants' moral commitment to non-prejudice, we ran moderation analyses for a multicategorical IV (Hayes, 2018). The analysis involved two dummy variables as independent variables: D1 (1 = loss and

0 = control and gain) and D2 (1 = gain and 0 = control and loss). The morality scales, namely, moral conviction, MID-internalization and MID-symbolization were tested as moderators, each in a separate model. Variables were not z-standardized or centered. As indicated before, we controlled for order of study materials (as covariate) in all three moderation analyses.

In the analysis where moral conviction was tested as a moderator, the two-way interaction between D1 (loss vs. control) and moral conviction on confronting intentions was not significant (*p* = .057; see Figure 8. See Table 5a for statistics for the interactions as well as the conditional main effects of both dummy variables and morality scales. Simple effects revealed that the loss mindset affected only those who were high on moral conviction (1 SD above the mean), such that it increased confronting intentions (*M* = 7.99) compared to the control condition (*M* = 7.48), *b* = 0.52, *SE* = 0.25, *t* = 2.11, *p* = .035, 95% CI [0.04, 1.00]. For those weakly morally convicted (1 SD below mean), this effect was not significant (*p* > .25).

Figure 8. Interaction between mindset framing condition (loss vs. gain vs. control) and participants' moral conviction on confronting intentions in Study 4 (on a 9-point scale).



The MID symbolization and internalization subscale did not moderate the relationship between loss (vs. control) mindset and confronting (*p*'s > .25), and simple effects were also not significant (*p*'s > .25). Additionally, there were no significant two-way interactions between D2 (gain vs. control) and any of the moral commitment scales (*p*'s > .25; And simple effects were also non-significant, *p*'s > .25). See Table 5b for estimated conditional means and simple effects for confronting intentions as a factor of condition and all moral commitment to non-prejudice scales.

Table 5a. The effect of moral mindset condition (control, loss, gain) on confronting intentions as a factor of moral commitment to non-prejudiced scales (controlling for order) in Study 4.

			Confronting intentions		
Moderator	*Predictor*	*B (SE)*	*t*	*p-value*	*95% CI*
Moral conviction					
	Moral conviction	.30 (.08)	3.80	.00	.15; .46
	D1 (Control vs. Loss)	-1.34 (.82)	-1.63	.10	-2.96; .28
	D2 (Control vs. Gain)	-.56 (.77)	-0.73	.47	-2.06; .94
	D1 x Moral conviction	.21 (.11)	1.91	.06	-.01; .42
	D2 x Moral conviction	.09 (.10)	0.85	.39	-.11; .28
	Order	.06 (.16)	0.37	.71	-.25; .37
MID-symbol					
	MID-symbol	.24 (.06)	4.02	.00	.12; .36
	D1 (Control vs. Loss)	.41 (.50)	.82	.41	-.57; 1.39
	D2 (Control vs. Gain)	-.06 (.47)	-.12	.91	-.97; .86
	D1 x MID-symbol	-.09 (.09)	-1.04	.30	-.26; .08
	D2 x MID-symbol	-.01 (.08)	-.11	.91	-.17; .15
	Order	.02 (.17)	.10	.92	-.31; .34
MID-internal					
	MID-internal	.27 (.07)	3.77	.00	.13; .42
	D1 (Control vs. Loss)	.17 (.86)	.20	.84	-1.51; 1.86
	D2 (Control vs. Gain)	.42 (.80)	.52	.60	-1.16; 2.00
	D1 x MID-internal	-.02 (.11)	-.20	.84	-.24; .20
	D2 x MID-internal	-.07 (.11)	-.62	.54	-.28; .14
	Order	-.05 (.17)	-.29	.77	-.38; .28

Table 5b. Simple effects and estimated conditional means for confronting intentions (9-point scale) in Study 4.

	Low on moderator (–1 SD)					High on moderator (+1 SD)				
	Control	Loss	Gain	D1	D2	Control	Loss	Gain	D1	D2
Moral convict.	6.43	6.23	6.34	*b* = -.20 *SE* = .27 *t* = -.74 *p* = .46 [-.73; .33]	*b* = -.09 *SE* = .26 *t* = -.33 *p* = .74 [-.59; .42]	7.48	7.99	7.69	*b* = .52 *SE* = .25 *t* = 2.11 *p* = .04 [.036; 1.00]	*b* = .21 *SE* = .24 *t* = .90 *p* = .37 [-.25; .68]
MID-symbol	6.54	6.68	6.46	*b* = .14 *SE* = .27 *t* = .51 *p* = .61 [-.40; .67]	*b* = -.08 *SE* = .25 *t* = -.33 *p* = .74 [-.58; .41]	7.61	7.36	7.48	*b* = -.26 *SE* = .26 *t* = -.99 *p* = .32 [-.76; .25]	*b* = -.12 *SE* = .27 *t* = -.47 *p* = .64 [-.65; .40]
MID-internal	6.57	6.62	6.62	*b* = .05 *SE* = .27 *t* = .18 *p* = .86 [-.49; .59]	*b* = .04 *SE* = .26 *t* = .17 *p* = .87 [-.46; .55]	7.49	7.47	7.31	*b* = -.03 *SE* = .26 *t* = -.10 *p* = .92 [-.53; .48]	*b* = -.18 *SE* = .26 *t* = -.69 *p* = .49 [-.70; .33]

Discussion

In Study 4, we found partial support for the idea that a moral mindset induction can increase levels of confronting racism. Consistent with our prediction (H1), we found that the moral loss framing, compared to a control, triggered more willingness to confront racism among those who were high on moral commitment to non-prejudice, specifically on moral conviction. There was no significant effect among those who were weakly convicted. The interaction, nevertheless, was not significant, *p* = .057, and these effects occurred with only one out of three potential moderators (the MID subscales did not have an influence on the relationship between loss mindset and confronting).

Furthermore, we failed to find evidence for our second prediction (H2) regarding gain mindset. We found that confronting rate in the moral gain condition did not significantly differ from the control condition, at any level of moral conviction or MID-symbolization or MID-internalization. It could be the case that the gain manipulation was not effective because, overall, our moral mindset priming was subtle and perhaps not sufficiently persuasive. Furthermore, another limitation was that we could not tell whether participants internalized the loss or gain messages of the vignettes, thus the current stimulus likely served simply at best as priming or nudging moral concerns, not as proper "conditioning".

In the next study, we created and tested an engaging and more "intrusive" moral mindset intervention. We are also not sure whether participants perceived the vignette situations as depicting prejudice given that we did not pre-test these vignettes for perceived prejudice. Additionally, we measured (hypothetical) willingness to confront racism and not actual behavior. Once participants need to make an (allegedly) real decision to confront, a different pattern of results may emerge (Crosby & Wilson, 2015). To overcome these limitations of the confronting measure, in Study 5, we employed a behavioral test of confronting racism, where participants believed that they were actually witnessing blatant prejudice and had an opportunity to contest it.

Study 5: Intervention study

Study 5 involved an online intervention we designed to induce a moral loss vs. gain moral mindset. We used the self-developed behavioral paradigm to measure actual confronting behavior.[31] Participants first filled out the scales of moral commitment to non-prejudice (same as those in Study 4). Then they were randomly assigned either to a moral loss or moral gain mindset intervention, or to an empty control condition. After a couple of days, we approached the same participants with a study allegedly pre-testing a behavioral economics game (which actually included the

[31] Same Trust game paradigm that was used in Study 1-2, reported in Chapter 2.

confronting measure). Participants believed that they were observing a game involving other participants. During the game, they witnessed a player being prejudiced and discriminating against an outgroup member (a Muslim individual) and had an opportunity to respond and thereby confront the racist player. Testing actual confronting allowed us to potentially capture real-life behavior.

Method

Participants and Procedure. We recruited 450 U.S. residents (North-Americans) through mTurk to complete the first part of the study for monetary compensation ($1.50). We ended up with two additional respondents, which is not unusual with mTurk. A-priori power analysis (G*Power 3.1; see Faul et al., 2009) for logistic regression (probabilities set to 0.25 and 0.15; R-squared other X = .50^1) revealed that a sample size of 247 is needed to achieve sufficient power of .80. Considering attrition and the exclusion criteria (attention and suspicion check), we collected 450 participants in the first part. In the survey, participants first completed the morality scales, then they were randomly assigned to a moral loss or moral gain mindset induction, or empty control condition (in which they completed the morality scales but were not exposed to any mindset related stimuli). This part of the study ended with demographic questions (*age, gender, SES, education level, liberal-conservative orientations, race/ethnicity* and *religion*. See Appendix E for full demographics).

Two days later, all respondents received a notification email advertising a new study (allegedly pre-testing a behavioral economics game). This email came from a different mTurk account in order to disguise that the studies were connected.[32] Only two respondents who identified as Muslim in the first part of the study were not invited back to due to ethical considerations (risk of psychological harm). Following our

[32] We contacted participants and collected data through an mTurk extension website called TurkPrime (Litman, Robinson, & Abberbock, 2016). We as researchers did not handle participants' identifying information such as their email address, we solely used their anonymized ID's.

email advertising the new study, we left the study open for 5 days to collect responses. With a 34% dropout rate, 297 respondents returned and went through the confronting measure. Data of participants who failed the attention check in the first part (same 1 item as in Study 4, n = 27, 9.1% of sample) or expressed suspicion about the game stimuli[33] (n = 10, 3.4% of sample) were excluded from data analyses, leaving 260 participants in the study (n = 82 in loss, n = 78 in gain, n = 100 in control condition[34]; 44.6% female, M_{age} = 36.08, SD_{age} = 10.27, range: 20-72).[35]

After completing the study, we messaged all participants for debriefing. We revealed to them the study purpose (how people react to uncomfortable intergroup situations, such as witnessing racism), the deception (that no racism had occurred as the game was pre-programmed), we reassured them the study was anonymous, and we provided them with our email address for further assistance.

Materials and measures. We included the same morality scales, moral conviction (4 items, α = .94), MID internalization (5 items, α = .84) and MID symbolization scales (4 items, α = .89) as in Study 4. The rest of the materials are described below.

Intervention stimuli. In the loss and gain conditions, participants were told that they would be asked to complete three tasks (see Appendix A for full material). Each of the tasks was aimed to induce a loss or gain mindset with respect to failure to confront immoral behavior (one task was general, the other two were specific to racism). In the first task a poster

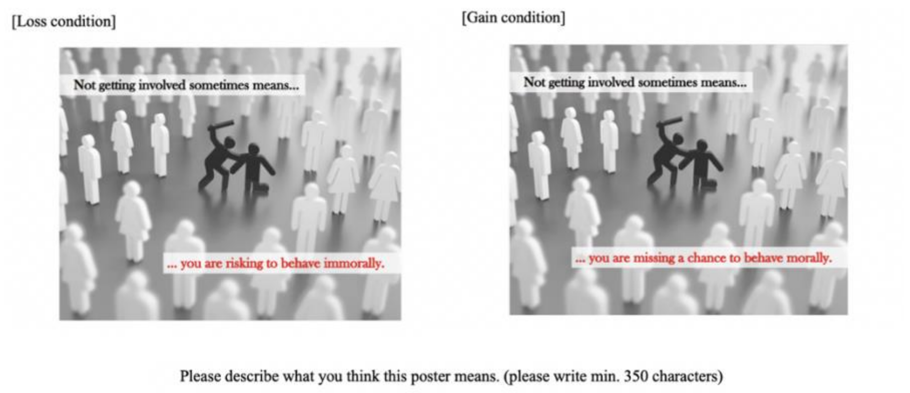
[33] We screened for suspicion about the realness of the game based on participants' messages to the alleged player they were observing (whether the participant asked or stated if the player is "real"/"bot") and based on an open-ended question at the end of study ("Please feel free to leave any comment/s or remark/s you may have."; whether participants wrote that the game was fabricated and/or that the racist remark was the actual aim of the study). Note that no participant made a comment that would lead us to believe that they figured out that the two parts were connected.

[34] There are more participants in the empty control condition than in the loss and gain conditions most likely because more respondents started these conditions without finishing it than respondents in the control (which was a shorter survey), but Qualtrics still calculated it toward equal randomization.

[35] *Socio-economic status* (1=destitute to 6=wealthy): *M*=3.45, *SD*=0.96, destitute (1.2%), poor (15.8%), so-so (32.7%), good (39.6%), better than most (8.8%), wealthy (1.9%). *Education*: less than high school (0.4%), high school diploma (33.5%), Bachelor's degree (49.6%), Master's degree (13.8%), PhD (0.8%), Other (1.9%).

appeared depicting a bystander situation (someone being physically attacked while others around do nothing) with a text either framed as moral loss (“Not getting involved sometimes means you are risking to behave immorally”) vs. gain ("Not getting involved sometimes means you are missing a chance to behave morally”), based on condition (see Figure 9). Participants were asked to write what they thought the poster meant. In the next task, participants were shown a video of a real event, depicting a British woman insulting immigrants on a bus, and then a text described a bystander passenger, who later allegedly reported his regrets of not confronting. Participants were asked to give a short account of their thoughts and feelings while imagining they are this passenger. The provided text box started with a stem sentence, which was framed according to condition (moral loss: “I feel like not intervening revealed a bad side of me…” vs. moral gain: “I feel like intervening would have revealed a good side of me…”). In the last task, a text described a Holocaust rescuer and an alleged account he gave about his actions, which was framed according to condition (moral loss: “…He once noted that not doing what he did would have cost him his moral virtue and he would have felt like a bad person…” vs. moral gain: “…He once noted that through this action he gained moral virtue and he feels he became a better person for doing it….”). Then participants were asked to describe this person’s potential thoughts and feelings about his own behavior.

Figure 9. An example task in the mindset intervention


[Loss condition]
Not getting involved sometimes means...
... you are risking to behave immorally.
[Gain condition]
Not getting involved sometimes means...
... you are missing a chance to behave morally.

Please describe what you think this poster means. (please write min. 350 characters)

Similar to Study 4, for a manipulation check, on a separate page we asked participants in the two experimental conditions the following questions: “What is suggested in the previous tasks about people's feeling and morality if they do not intervene in those situations? (pick the convenient sentence starter and continue the sentence)”. Answer options were 1 = “they miss a chance to gain…. (text entry)”, or 2 = “They risk to lose… (text entry)”. The chosen option enabled us to assess whether we succeeded to prime a gain or loss mindset (we did not analyze how they completed the sentence).

Confronting stimuli. We used (and minimally shortened) the “Trust game” paradigm from Study 1-2 (see Chapter 2). We used and altered the “trust game” (Berg et al., 1995; Charness & Rabin, 2002), where Player A decides how much money out of an initial endowment to send to another subject, Player B. The sent amount is then multiplied by 3 and Player B decides how much of the money received to send back to Player A. To conceal the purpose of the study, participants were told that we are testing how observing influences trusting behavior and how does gender (of the players and observers) influence trusting behavior. Participants were given instructions on the game (see Appendix A), and they were “trained” on the rules. They were told that they would first be assigned to a player and observe his/her rounds and only after they would play this game themselves for money. Participants were further explained that this player they observe could initiate private messaging with them. Allegedly this was enabled in order for the “observer” to feel more real to the player, in reality this was done in order to manipulate racism and enable confronting. In order to strengthen our cover story, we asked participants not to share the purpose of the study with the players while they were messaging.

Participants then entered a different site to observe the game (in reality the game observed was pre-programmed). To make participants feel present in the situation we asked them to provide their nickname as well (they appeared with this throughout the game). All participants were assigned to observe a player called Mark, and then observed two decoy rounds and exchanged some messages with Mark (in these messages Mark addressed the participant by their nickname and his responses were written to fit any message the participant replied with). In the first observed round Mark played against Kip and gave half of his money, then against Nica and gave all of his money (both partners returned the money fairly). Then, Mark's partner appeared as Hakim (a Muslim name), and to him Mark decided to give no money. Then, Mark privately messaged the participant saying: "You can't trust those damn Muslims" (see Figure 5 for scenes from the game). Participants thus witnessed a discriminatory act and an explicitly racist comment. Beneath the message, participants had a chance to either press 'continue game' or 'reply'. Then, for all participants a message appeared on the screen indicating that there was a problem registered in the system, and the game terminated.

Confronting. Participants who chose to continue the game, or to reply but left the message box empty, or those who responded in a non-confronting way - were all labeled as 'not confronting' and coded as '0'. Responses that questioned or reproached the player for his behavior and statement were labeled as 'confronting' and coded as '1'. Responses that are unclear as to their intentions (confronting or not) were coded as other and treated as missing values in the main analyses (as indicated in the participants section part, those who communicated suspicion here about the study were excluded from analyses). These responses were coded by two authors of the manuscript blind to conditions, and disagreements (n = 3) were discussed.

Results

Preliminary analysis. First, we tested and found no significant differences across conditions on the morality scales, *p*'s > .17, or on demographics, *p*'s > .22 (see Table 6 for means, standard deviations and correlations between study variables). We found that the mindset manipulation was successful in communicating the sense of moral loss vs. gain, χ^2 (1) = 37.95, *p* < .001, Cramer's V = 0.49. Specifically, significantly more people indicated the loss (vs. gain) response in the loss condition

(76.8%), and significantly more participants indicated the gain (vs. loss) response in the gain condition (71.8%).

Next, we analyzed the responses of the confronting message. Considering all conditions, we found that 120 participants pressed to continue the game (were thus coded as not confronting, ‘0’), 140 pressed to reply. Among repliers, 108 participants (41.5% of all participants) confronted the racist perpetrator (coded as confronting, ‘1’), 19 people wrote messages that expressed consent or simple acknowledgment (e.g., “OK”) (they were all coded as not confronting, ‘0’), and 13 responses were ambiguous and thus coded as ‘other’ and were not used for data analyses.

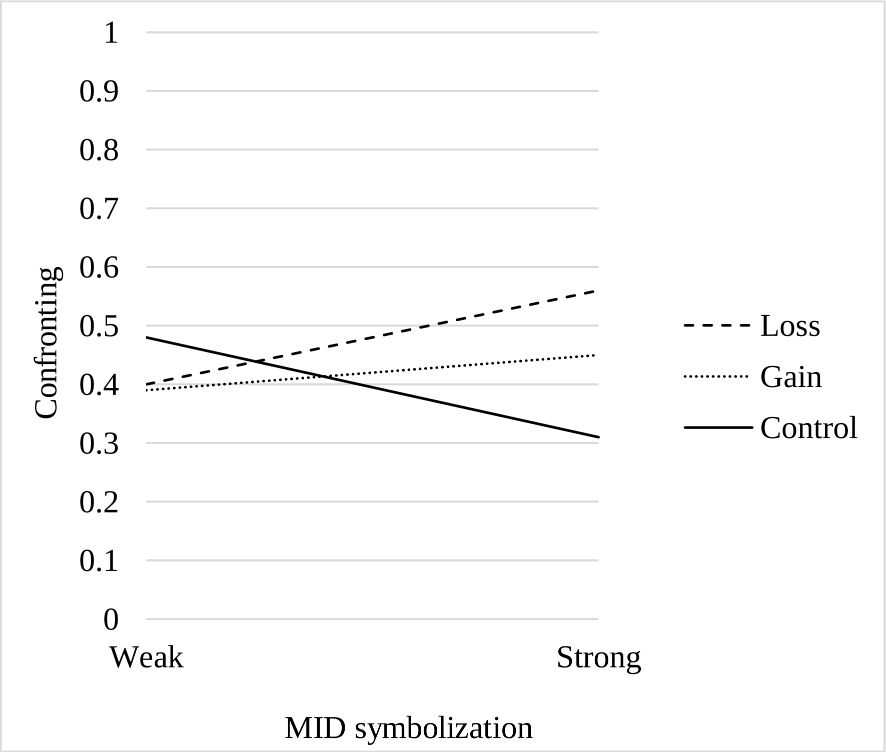
Table 6. Percentages or means and standard deviations (M and SD) and correlations between study variables in Study 5.

	M (SD)	Conf. (0: not, 1: yes)	MC	MID int.	MID symb	Cons. -Lib.	SES	Edu
Confronting (0: not, 1: yes)	*56.3%, 43.7%*	-						
Moral conviction	*7.39 (1.67)*	.16*	-					
MID-internal	*7.33 (1.68)*	.21**	.52**	-				
MID-symbol	*5.87 (1.97)*	–.03	.31**	.19**	-			
Conservative –Liberal	*6.38 (3.11)*	.11	.24**	.14**	–.01	-		
SES	*3.45 (0.96)*	-.12	-.02	–.25**	.19**	-.10	-	
Education	*2.81 (.70)*	-.06	-.07	–.18**	.06	-.05	.34**	-
Age	*36.08 (10.27)*	.15*	.11	.22**	–.01	-.08	-.07	.10

Note. * p < .05, ** p < .01. Moral conviction (MC) and MID scales were on a 9-point continuous scale. Conservative-liberal dimension was on a continuous slider from 0 (conservative) to 10 (liberal). SES was on 6-point and education was on a 5-point ordinal scales.

Main analyses. Analysis strategy was the same as in Study 4, except this time we ran logistic regression because our DV was dichotomous (confronting or not). As indicated in Table 7a, there was a significant two-way interaction between D1 (loss vs. control) and MID-symbolization on confronting intentions. A simple-effects analysis revealed that the loss mindset affected only those who were high on MID-symbolization (1 SD above the mean), such that it increased confronting intentions (prob. = 0.56, odds = 1.27) compared to the control condition (prob. = 0.30, odds = .43), b = 1.10, SE = .46, Z = 2.37, OR = .34, p = .018, 95% CI [.19; 2.01] (see Figure 10). This effect was not significant among those weakly committed (1 SD below mean; p > .25).

Figure 10. Interaction between mindset framing condition (loss vs. gain vs. control) and participants' MID-symbolization on confronting in Study 5 (yes or no; visualizing probabilities).



Other morality scales (MID internalization and moral conviction) did not moderate the relationship between loss (vs. control) and confronting (*p*’s > .25), and simple effects were also not significant (*p*’s > .24). Additionally, there were no significant two-way interactions between D2 (gain vs. control) and any of the moral commitment scales (*p*’s > .10), nor significant simple effects (*p*’s > .14). (See Table 7b for probabilities and simple effects for confronting as a factor of condition and moral commitment to non-prejudice scales.)

Table 7a. The effect of moral mindset condition (loss, gain, control) on confronting behavior (0: not confronting, 1: confronting) as a factor of the moral commitment to non-prejudiced scales in Study 5.

		Confronting behavior			
Moderator	*Predictor*	*B (SE)*	*Z*	*p-value*	*95% CI*
Moral conviction					
	Moral conviction	.21 (.15)	1.36	.18	-.09; .51
	D1 (Control vs. Loss)	-.38 (1.66)	-.23	.82	-3.63; 2.88
	D2 (Control vs. Gain)	.66 (1.59)	.42	.68	-2.44; 3.77
	D1 x Moral conviction	.10 (.22)	.47	.64	-.32; .53
	D2 x Moral conviction	-.08 (.21)	-.37	.71	-.48; .33
MID-symbol					
	MID-symbol	-.20 (.10)	-2.01	.05	-.39; -.01
	D1 (Control vs. Loss)	-1.76 (1.02)	-1.72	.09	-3.76; .25
	D2 (Control vs. Gain)	-1.43 (1.01)	-1.41	.16	-3.41; .56
	D1 x MID-symbol	.37 (.17)	2.18	.03	.04; .70
	D2 x MID-symbol	.27 (.16)	1.65	.10	-.05; .59
MID-internal					
	MID-internal	.23 (.14)	1.65	.10	-.04; .51
	D1 (Control vs. Loss)	.02 (1.58)	.01	.99	-3.08; 3.12
	D2 (Control vs. Gain)	-.75 (1.64)	-.45	.65	-3.97; 2.47
	D1 x MID-internal	.05 (.21)	.23	.82	-.36; .45
	D2 x MID-internal	.11 (.21)	.54	.59	-.30; .53

Table 7b. Probabilities (odds in brackets) for each condition and simple effect statistics for confronting action (0 = didn't confront, 1 = confronted) in Study 5.

	Low on moderator (–1 SD)					High on moderator (+1 SD)				
	Control	Loss	Gain	D1 (Control vs. Loss)	D2 (Control vs. Gain)	Control	Loss	Gain	D1 (Control vs. Loss)	D2 (Control vs. Gain)
MC	.32 (.47)	.37 (.59)	.37 (.59)	*b* = .21 *SE* = .50 *Z* = .42 *p* = .68 [-.77; 1.18] *OR* = .80	*b* = .23 *SE* = .49 *Z* = .46 *p* = .64 [-.73; 1.18] *OR* = .80	.48 (.92)	.61 (1.56)	.48 (.92)	*b* = .53 *SE* = .45 *Z* = 1.18 *p* = .24 [-.35; 1.42] *OR* = .59	*b* = -.02 *SE* = .43 *Z* = -.05 *p* = .96 [-.87; .82] *OR* = 1
MID sym.	.48 (.92)	.40 (.67)	.39 (.64)	*b* = -.34 *SE* = .45 *Z* = -.75 *p* = .45 [-1.23; .55] *OR* = 1.37	*b* = -.39 *SE* = .46 *Z* = -.84 *p* = .40 [-1.28; .51] *OR* = 1.44	.30 (.43)	.56 (1.27)	.46 (.85)	*b* = 1.10 *SE* = .46, Z = 2.37 *p* = .02 [.19, 2.01] *OR* = .34	*b* = .67 *SE* = .45 *Z* = 1.49 *p* = .14 [-.21; 1.55] *OR* = .51
MID int.	.31 (.45)	.38 (.61)	.29 (.41)	*b* = .29 *SE* = .49 *Z* = .60 *p* = .55 [-.66; 1.25] *OR* = .74	*b* = -.10 *SE* = .51 *Z* = -.19 *p* = .85 [-1.10; .91] *OR* = 1.10	.49 (.96)	.60 (1.5)	.56 (1.27)	*b* = .45 *SE* = .44 *Z* = 1.01 *p* = .31 [-.42; 1.31] *OR* = .64	*b* = .27 *SE* = .43 *Z* = .63 *p* = .53 [-.58; 1.13] *OR* = .76

Discussion

In Study 5, we developed and tested an online moral mindset intervention and a couple of days later employed an online behavioral paradigm to measure real acts of confronting racism. We found partial support for our prediction (H1) that a moral loss mindset can increase confronting of racism among those high in moral commitment to non-prejudice. Nevertheless, also in Study 5 the effect emerged only with one out of the three considered moderators, and, unlike in Study 4, in this study a MID subscale (symbolization) moderated the effect of moral loss mindset on confronting (and not moral conviction). We discuss potential reasons for this inconsistency in detail in the general discussion. As in Study 4, confronting rate was not significantly affected by the gain mindset (vs. control) at any level of moral commitment. We address the lack of support for our expectation in this respect (H2) in the general discussion.

Interestingly, MID-symbolization seemed to have a negative relationship with confronting. Specifically, in the control group, confronting tended to be higher among those low in MID-symbolization compared to those higher on this scale. Note that this scale asked about the respondents' behavioral commitment to non-prejudice in everyday life, about their hobbies, activities, memberships in organizations that reflect these values. Those high on symbolization are usually driven to publicly exhibit their moral self, and they are motivated by recognition and reputational gains from engagement in moral deeds (Winterich et al., 2013; Winterich et al., 2013). It could be the case that participants high on this measure felt excused from confronting because their public activities provided them with moral credentials (see Monin & Miller, 2001). This might have not been the case for those who had no such credentials (i.e., low on MID-symbolization). Thus, what the loss mindset intervention possibly did, is that it induced individuals high on MID-symbolization to think about losing these moral credits, thereby motivating their confronting. Put differently, it is possible that the moral loss mindset acted as a buffer to this general moral licensing process.

Finally, we should point out that our manipulation check, assessing the extent to which we succeeded to communicate a loss/gain mindset was limited because it did not involve a comparison with the control condition. In the manipulation check we asked: "What is suggested in the previous tasks about people's feeling and morality if they do not intervene in those situations? ". Answer options were either "they miss a chance to gain…. (text entry)", or "they risk to lose… (text entry)". This manipulation check question would have not made sense to participants in the empty control condition because they were not exposed to any tasks.

General discussion

People who intervene in times of racial and ethnic discrimination often describe their actions as driven by a need to avoid a sense of moral failure. Holocaust rescuers often explained their decision to help along the line of "Can I live with myself if I say no?" (Fogelman, n.d.). Similarly, in the example of protest against deportation, Ersson herself said in an interview, “I knew that I couldn’t back down because it was my name that was on the ticket. I had to do what I could”. In the present research, we found evidence that this sense of moral failure can indeed motivate people to confront racism. Across two studies, we tested the effects of thinking about moral loss and moral gain on contesting racism in light of people’s non-prejudice moral commitment. In Study 4, participants were presented with vignettes depicting racist scenarios, where we varied the description of potential moral concerns (loss vs. gain vs. control) and assessed participants’ self-reported intentions to confront the racist act. In Study 5, participants went through a moral mindset intervention (that was intended to induce a loss or gain mindset), or empty control, and a few days later, we employed a behavioral paradigm to measure real action to confront racism.

The studies provide partial support to our predictions. First, we failed to find support for the hypothesis regarding the moral gain mindset (H2). This mindset did not seem to be effective in increasing confronting rate in comparison to the control group at any level of moral commitment to non-prejudice. Regarding our other hypothesis (H1), we predicted and found partial evidence that moral framing can affect the tendency to confront racism, and this is dependent on participants’ non-prejudiced moral commitment. Across studies, among those with high moral commitment to non-prejudice, a loss mindset led to more confronting, compared to the control condition (H1). Likely, the loss framing activated motivation to safeguard one’s moral non-prejudiced self-concept (Dutton & Lennox, 1974; Monteith, 1993). However, in each of the studies a different variable moderated the effect. In Study 4, the moral conviction about prejudice scale (adapted from Skitka & Morgan, 2014) moderated the effect of loss mindset (vs. control) on confronting (although the interaction was not significant, only the relevant simple effect), while in Study 5 it was the symbolization facet of the moral identity-prejudice scale (adapted from Aquino & Reed, 2002) that significantly moderated these effects. We employed these scales, and we did not have specific prediction as to (when and) which moral commitment construct would have influence on the relationship between moral gain/loss mindsets and confronting.

Due to these different moderation effects, our study findings may be incidental (and reflect false positive results), and thus attempts of replication are advisable in the future. However, there were notable differences between the two studies that may account for the differing effects of the morality scales. Namely, the manipulation in Study 4 involved an imagined scenario that focused on the victim of prejudice, and the participant was not the actual person in the described situation who had the responsibility to confront. In Study 5, the prejudiced situation was perceived as real (and not hypothetical), making the participants believe they had to make an actual choice to confront or not, rendering less focus on the harm done to the victims and more focus on the responsibility and actions of the participant. Correspondingly, those high on MID-symbolization are usually driven to publicly exhibit their moral self, and they are motivated by recognition and reputational gains from engagement in moral behavior (Schaumberg & Wiltermuth, 2014; Winterich et al., 2013a; Winterich et al., 2013b). Thus, in Study 5, it was those high on MID-symbolization who became encouraged to confront, perhaps because they felt personally involved in the (allegedly) real-life situation and under loss induction they felt that their own moral public identity and reputation is at risk. Even more so if they thought that others may see their actions such as the perpetrator, other players, or the experimenter.

At the same time, compared to MID-symbolization, moral conviction is a relatively other-oriented moral attitude given that it reflects internally entrenched beliefs (Skitka, 2010; Skitka, 2014; Skitka et al., 2005), which likely renders more focus on the actual subject of this strongly held belief, in our case on the target of prejudice. Accordingly, we speculate that in Study 4 where there was a stronger focus on the victims of prejudice, the moral loss induction could trigger specifically those high on moral conviction to confront. Having said that, to our knowledge no prior research have tested or discussed these moral constructs in the same work, therefore our outlined theoretical distinction was speculative**.** The current explanations to why different morality constructs pertained more to confronting under differing situational cues warrants further investigation as in the current research we are unable to answer that.

While the empirical evidence to our proposed effect is limited, we offer an important initial step toward investigating the understudied effect of anticipated moral cost on not confronting prejudice. Our findings partially align with research on regulatory focus, which indicates that individuals under prevention focus (corresponding to the loss mindset) are more likely to engage in action aimed at amending injustice directed towards their own group, than those in a control group, and this is not the case for those under promotion focus (corresponding to a gain mindset, e.g., Sassenberg & Hansen, 2007; Zaal et al., 2012). This effect is more pronounced if individuals hold a strong moral conviction about the fair treatment of their group (Zaal et al., 2011). We found consistent pattern in the domain of morality in the context of third-party intervention, showing that those induced to think in terms of a loss to their morality, were more likely to confront that those in a control condition, if they were committed to non-prejudice.

For people with promotion focus, taking action depends heavily on the (perceived) instrumentality of the action, i.e., on the expectation of success (Shah & Higgins, 1997). Therefore, trying to motivate action through reframing the action's moral goal in promotion-oriented terms would only be effective when the likelihood that the action will succeed is high (Quinn & Olson, 2011; Zaal et al., 2012). In our studies, the way we framed the moral gain mindset did suggest that if the individual does confront, he or she would likely succeed in gaining a positive and moral self-regard. At the same time, we did not (necessarily) communicate that confronting will be successful in for example, changing the perpetrator's mind or in helping the victim. This is possibly why we did not find the gain mindset affecting confrontation of racism.

Our findings also reflect, to some extent, the loss aversion effect, which states that losses inflict psychological harm to a greater degree than gains gratify (Tversky & Kahneman, 1991; 1992). However, when people witness racism and contemplate whether to confront, the potential moral self-concept loss may not actually (psychologically) *equal* the potential gain. Prospect theory was, for the most part, applied to constructs (such as monetary investment) that can be readily quantified. Individuals' relation to their own moral self-concept is not necessarily the same as relations to their material possessions. Thus, the extent to which we can apply loss aversion theory to the current intergroup context is debatable.

Our findings however cannot be explained by a moral priming effect, whereby activating certain aspects of morality in memory (e.g., a just prototype; Osswald et al., 2010) increases morally courageous behavior. For one, in both studies, all participants (including those in the control group) responded to scales that were explicitly about morality. Secondly, in Study 4, participants in the control group read the same vignettes that included the prejudicial situation, the opportunity and the pro and con concerns of intervening (while moral gain and loss arguments were not mentioned). Finally, our results showed that the gain mindset manipulation did not influence confronting, and moral commitment moderated only the effects for the loss mindset – rendering it unlikely that priming, or experimental demands can explain our findings.

In general, when it comes to the question of witnessing racism, lay and empirical discussion is usually focused on the personal costs of confronting. Namely, on people's courage to stand up against injustice despite the anticipation of substantial costs to themselves. Such sacrifices are without question admirable and should be recognized. Nevertheless, not much is being said about the personal benefits of confronting, or more correctly, about the personal moral costs of *not* confronting. The present work sheds light on this perspective, by showing that when one cares about being non-prejudiced, the potential loss of one's sense of morality if action is not taken can actually trigger confronting behavior.

Given that confronting in our studies was influenced by the person's consideration about their own morality, and not only about standing up for the victims, can we still consider it a morally courageous behavior? This resonates with the age-old question about the nature and existence of selfless good deeds (Kant, 1785; Nietzsche, 1878), and whether if an individual benefits from their prosocial behavior is that act ultimately egoistic (self-oriented; Andreoni, 1990) or it may nonetheless be considered courageous and altruistic (other-oriented; Batson, 2011; Batson & Shaw, 1991). This remains an unanswered philosophical question. However, considering the motivation of Holocaust rescuers and of Ersson that was mentioned before, we believe that a person's concern about their own morality, triggered by the treatment of *another group* is still at some level an other-oriented concern, that could benefit victims, and mitigate bias among perpetrators. Thus, when considering the tangible outcomes of confronting, we see it as socially beneficial, even if the motive was egoistic.

Limitations and future directions

Following the previous argument, one could also question whether confrontation in our studies were morally courageous in the sense of involving personal costs to participants. In Study 4, in both imagined scenarios we explicitly stated the personal costs involved to confronting (e.g., jeopardize your position and respect at work; be verbally or physically attacked). In Study 5, in an (allegedly) real online situation participants were likely concerned that if they confront, they might "lose face", or the perpetrator may reply aggressively, or they sabotage the game and will not get their money. Note, that one study tested a hypothetical situation and the other one was online, thus the generalizability of the findings are limited in this respect. Future research where the study predictions are tested in an in-person, offline context is needed.

In a similar vein regarding external validity, it is a question whether we can generalize our findings to other countries because both of our studies were conducted with U.S.-based participants (while based on the variance in demographics it was a diverse sample).[36] Our findings may not

[36] Additionally, mTurk samples are considered being close to representative to the general population.

generalize to countries with different social norms about expressing racism.

Furthermore, the beneficial effect of induced loss mindset intervention was limited to those who were morally committed to non-prejudice. However, this is an important population to consider in encouraging for intervening because our findings reported in Chapter 2 show that those with non-prejudiced ideas (perceive confronting prejudice as important) are especially likely to justify their inaction in face of racism through actually derogating the outgroup. It remains a direction for future research to conceive messages (most likely outside of the moral mindset domain) that would motivate confronting among those who are less committed to non-prejudice.

Another gap in our research is that we did not measure people's innate or automatic moral mindset when witnessing racism and how that motivates confronting. We do not know whether people who are (or are not) morally committed to non-prejudice have a stronger inherent tendency to take a moral loss (or gain) perspective during contemplation of confronting racism. It is also not clear how this natural tendency interacted with our induction of varying moral mindsets. We did find that the moral ought vs. ideal orientation did not play a role in the tested mechanisms, however this measure was not specific to prejudice, therefore future research is still needed to investigate inherent motivations. Moreover, regarding the operationalization of moral loss and gain, it remains a question whether participants internalized these actually quite abstract moral messages we aimed to manipulate, as we did not find a way to pre-test it.

Practical implication

Messages that promote confronting are important because racism is widespread, and while confronting can be effective in decreasing prejudice in perpetrators and bystanders (e.g., Czopp & Monteith, 2003; Czopp et al., 2006), people who witness racism rarely intervene, although they believe they would (e.g., Crosby & Wilson, 2015; Karmali et al., 2017; Kawakami et al., 2009). In the current study, we identified a process and defined messages that can be utilized as a potentially effective intervention tool to increase (some) people’s tendency to confront racism, for example in the form of social ad campaigns or workshops. Notably, in Study 5, the effect of our intervention seemed to endure across days, and to affect actual confronting behavior. Additionally, due to the feasibility in utilizing the confronting measure, it can be useful for assessment in devising and testing similar interventions in the future.

Conclusion

Moral courage is a willingness to take a stand in defense of one’s own moral principles even when others do not (Miller, 2000; Skitka, 2012). In this research, we tested a way to increase morally courageous behavior and motivate confronting intentions during a situation when people witness racism. We found that exposing people to messages about prospective personal moral failure in regard to not intervening was potentially effective in promoting speaking up against racism. While moral courage is often thought of as a solely altruistic act, we argue that the role of an individual’s consideration of their own morality should not be dismissed and can be used to the advantage in encouraging moral behavior in face of racist acts.

Chapter 4: Main discussion

Overview of the findings

While people generally believe they would stand up against prejudice and discrimination, historical precedents and empirical evidence suggest that bystanders often fail to do so (Crosby & Wilson, 2015; Karmali et al., 2017; Kawakami et al., 2009). Such inaction is harmful because confronting, especially if done by non-stigmatized individuals, can effectively change people's minds, reduce prejudice in the perpetrator (Czopp & Monteith, 2003; Drury & Kaiser, 2014; Gulker et al., 2013) and reaffirm egalitarian norms and standards in the surrounding social environment (Blanchard et al., 1994; Rasinski & Czopp, 2010; Schultz & Maddox, 2013). In the current research, across seven experiments, we investigated the consequences of witnessing and not confronting prejudice, and the potential psychological messages that would motivate (non-target) bystanders to speak up. For the purpose of the current research, to test actual confronting, I developed and used an online behavioral paradigm, where participants believes they witnessed a prejudice slur and discriminatory act against an outgroup and had an opportunity to confront the perpetrator.

In the first research (Chapter 2), we aimed to understand how those who witness prejudice and discrimination are impacted by such incidents and demonstrate a path through which prejudice perpetuates and intensifies over time. We draw on cognitive consistency theories to propose and test that people who witness prejudice and do not contest it (although having the opportunity to do so), subsequently endorse more negative intergroup attitudes to justify and reconcile their attitudes with their inaction. We conducted five experiments in two countries (N = 922), in the US and in Hungary, in various intergroup contexts where non-target participants witnessed prejudice directed at outgroup minority who was in the US either Black American (Pilot studies), Muslim (Study 2), or Latin American (Study 3), or Jewish in Hungary (Study 1). Across all studies, we used the online paradigm to test actual (non-)confronting behavior. Following two pilot studies, in Studies 1–3, using a mixed within- and between-subjects design, we assessed participants both prior and following witnessing of a prejudiced event (pre- and post-test). This design enabled us to test overtime changes in attitudes among those who did not confront, and to compare those changes to control groups, in order to show that people *come* to endorse more negative outgroup attitudes as a function of witnessing and not confronting prejudice. In Studies 1–2, in the control condition, participants observed another type of prejudice, not rooted in intergroup membership (but "interpersonal") and had an opportunity to react. We tested and predicted no change in attitudes among those who did not confront interpersonal bias. In Study 3, we added another control condition, where participants observed the same intergroup prejudice but did not have an opportunity to confront – they were merely exposed to prejudice.

According to our prediction, we found that those participants who witnessed intergroup prejudice and had an opportunity to confront the perpetrator, but did not do so, endorsed more negative outgroup attitudes relative to their own attitudes prior to the incident (Studies 1–3). Additionally, they showed more negative intergroup attitudes (increased outgroup prejudice, trivialization of the witnessed incident, and some responsibility denial for intervening) than those who witnessed and did not confront other (non-intergroup) type of bias (Studies 1-3), or those who witnessed the same intergroup prejudice scenario, but did not have a chance to confront (Study 3). For these control groups there was also no significant change in attitudes from prior to following the incident – unlike for those who witnessed and did not confront intergroup prejudice. We further found that the motivated prejudice effect did not occur among those who initially did not value confronting prejudice (Study 3; although the moderation was not significant), likely because inaction did not contradict their personal values, thus they did not seek justification for not confronting. This boundary condition, the effect of (outgroup) attitude change and especially trivialization and responsibility denials (which are typical dissonance-reduction strategies; McGrath, 2017), and the ruling out of mere exposure to prejudice as alternative explanation – all provides some indirect evidence to the dissonance-induced self-justification account.

In our second research (described in Chapter 3), across two experiments (N~710, conducted in the US), we investigated whether the prospect of moral loss (failure) or gain (success) relating to intervening can motivate people to confront prejudice. We considered that people's initial moral commitment to non-prejudice likely qualifies the effectiveness of the moral messages. Drawing on research on regulatory focus and loss aversion theories, we predicted that a moral loss framing/mindset would significantly increase confronting tendencies among those strongly morally committed to non-prejudice (possibly due to a desire to safeguard their moral self-concept), but not among those weakly committed. We also predicted that a moral gain (vs. control) framing/mindset would drive confronting among those who are weakly committed to non-prejudice (possibly to enhance their moral self-concept) and would not affect those strongly committed. We conducted our studies in the US where the outgroup minority was Latin-American and/or Muslim-American. In Study 4, participants were presented with vignettes depicting prejudiced scenarios, where we varied the description of potential moral concerns (loss vs. gain vs. control) and assessed participants' self-reported intentions to confront the prejudiced incident. In Study 5, participants went through a moral mindset intervention (that was intended to induce a loss or gain mindset), or empty control, and a few days later, we employed our behavioral paradigm to measure real confronting action. Opposed to predictions, the moral gain framing/mindset did not affect confronting at any level of moral commitment to non-prejudice. However, we found evidence that certain moral messages can affect the tendency to confront prejudice. Across studies, among those with high moral commitment to non-prejudice, a loss mindset led to more confronting, compared to the control condition. Likely, the loss framing activated non-prejudiced individuals' motivation to safeguard their moral non-prejudiced self-concept (Dutton & Lennox, 1974; Monteith, 1993).

Theoretical and applied contributions

In regard to our first research on the motivated prejudice effect, our findings correspond to the literature, which showed that women who initially valued confronting and were given the opportunity to confront, but did not, made more favorable evaluations of the sexist perpetrator, compared to those who had no chance to confront, and also devalued confronting socially inappropriate behavior in general – possibly all in order to reduce dissonance for inaction (Rasinski et al., 2013). In the present research we proposed that such dissonance-justification mechanism can also occur among bystanders not belonging to the target's group. This shift in focus enabled us to go beyond evaluations of the perpetrator, to identify a devastating cycle of how prejudice potentially perpetuates and intensifies overtime. The impact of failing to confront prejudice on non-targets is scarce as prior work on non-targets' reaction focused on the discrepancy between anticipated and actual reactions to prejudice (such as apathy or lack of interpersonal rejection of the perpetrator, e.g., Karmali et al., 2017; Kawakami et al., 2009), or between actual and hypothetical levels of confronting (e.g., Crosby & Wilson, 2015). To the best of our knowledge, it has not been tested previously how bystanders' failure to confront actually amplifies their own prejudicial beliefs and leads to trivialization of the witnessed incident, and to denial of responsibility for intervening. Beyond advancing theoretical knowledge on the intrapersonal–intergroup ramification of witnessing prejudice, the present work sheds light on the practical importance of addressing bystanders' failure to confront prejudice and discrimination.

Based on the found motivated prejudice effect, we developed our second research to create moral messages about *prospective* intrapersonal costs that may motivate confronting. In this work, our findings align with research on regulatory focus, which indicates that (target) individuals with prevention focus (corresponding to the loss mindset) are more likely to engage in action aimed at amending injustice directed towards their own group and this is not the case for those under promotion focus (corresponding to a gain mindset, e.g., Sassenberg & Hansen, 2007; Zaal et al., 2012). This effect is more pronounced if individuals hold a strong moral conviction about the fair treatment of their group (Zaal et al., 2011). Our findings also reflect the loss aversion effect, which states that losses inflict psychological harm to a greater degree than gains gratify (Tversky & Kahneman, 1991; 1992). Contributing to this prior work, our findings offer an important step toward investigating the understudied effect of anticipated moral costs (of inaction) on motivating moral behavior, in this case, confronting prejudice as a non-target bystander. Similarly in the confronting prejudice literature, while the anticipated costs of *acting* are tested and discussed quite extensively (for a review, see Mallet & Monteith, 2019), bystanders' anticipated intrapersonal (moral) costs of *inaction* is much less investigated. The present work sheds light on this perspective by showing that when one cares about being non-prejudiced, the *prospect* of loss of one's sense of morality if action is not taken can actually motivate confronting behavior. On this note, it is interesting to consider how *actual* intrapersonal cost of inaction led to outgroup derogation, meanwhile *anticipated* intrapersonal cost of inaction led to outgroup help. As mentioned in the discussion of Chapter 2 it is possible that non-confronters saw no mode available for positive compensation (outgroup helping) at the moment to ease their conscience so instead they derogated. Those who were made to consider their conscience and then they were provided with a positive route, they chose to engage in that. Future research should investigate these potential mechanisms more in-depth.

In this research, we identified a process and defined messages (and tasks) that can be utilized as a potentially effective intervention tool to increase (some) people's tendency to confront prejudice (or perhaps even other forms of immoral behavior). Notably, the effect of the self-developed intervention seemed to endure across days, and to affect actual confronting behavior. These intervention messages could be implemented in the field in the form of social ad campaigns or workshops or used by civil organizations for developing tools for promoting tolerance, and applied in companies or schools, where the community is diverse, and instances of bias can readily occur and are likely to go uncontested. Developing viable and evidence-based interventions (Szekeres, 2020) are especially important given that there is scarce empirical (hard) evidence to what psychological messages or tools can promote confronting prejudice for non-target bystanders. One notable research in this area was conducted by Rattan and Dweck (2010) who found that among targets of prejudice, growth (vs. fixed) mindset of personality can motivate confronting (i.e., believing that with confronting one can change the views of the perpetrator). It is most likely that such messages would be effective for non-targets, as well, however to date this was not tested.

Lastly, I believe that a major strength of our research, and contribution to current knowledge in the field, is the employment of the online behavioral paradigm. This way we were able to place participants in an allegedly real situation, and measure actual confronting behavior. This allowed us to potentially generalize our findings to real-life situations. Furthermore, since the paradigm was online it allowed us to collect data among participants outside the laboratory, thereby easing the feasibility of conducting multiple studies in different intergroup contexts, and of gaining more diverse samples in each study (not only psychology students), and larger sample sizes – all which is not typical of prior studies in this topic. Additionally, by conducting the research online and providing participants with a true sense of anonymity, we also minimized the emotional obtrusiveness of an otherwise distressing situation. Based on our experience with the small-scale in-lab study that I conducted in the beginning of this dissertation research, it is important to acknowledge this ethical consideration, as well. Finally, there is also practical and applied benefits to developing and validating the online behavioral paradigm, as it is easy and feasible to implement it for devising and assessing the effectiveness of field interventions in the future (bystander or other prejudice reduction).

Limitations and future directions

While there are many benefits of conducting the research online, the external validity and so generalizability of our findings is limited in regard to confrontation that occurs in face-to-face interactions. There are some differences given an in-person experience, such that the situation may feel more shocking, responsibility to act more emphasized, less external justification for not acting (offline confronting may be perceived more effective), or perhaps more external justification (e.g., avoiding physical attack). Another possible (or evident) difference between offline and online contexts is that confronting rate is lower in the former (see Crosby & Wilson, 2015, who found 0% confronting) compared to, for example, what we found in our studies (in Hungary it was a “realistic” 8%, but in the US studies they were between 20-40%). Although we attempted to create costs for confronting, such as interpersonal (e.g., loosing face, or getting an aggressive reply from the perpetrator) and economic (e.g., the perpetrator will penalize confronters by not sharing money with them), people possibly anticipated more (intense) costs to confronting in offline situations, thus deterring them to confront more so than online. Yet, overall, we assume that both investigated phenomena, the motivated prejudice effect and the impact of the moral loss mindset, are driven by psychological processes that are not specific to online contexts and would also manifest in face-to-face situations. Thus, the used paradigm has relevance to naturalistic forms of social interactions. Not only people spend more and more time online, but also more interpersonal interactions are becoming virtual, and plenty of socio-political discussions and activism are carried out online (Pew Research Center, 2021). Nevertheless, in future research, also in order to strengthen ecological validity, the found effects should be replicated in face-to-face contexts.

Furthermore, in the behavioral paradigm, the perpetrator shared the prejudiced slurs *privately* and *only* with the participants (also making it clear that they observed discrimination) – rendering the bystanders solely responsible to speak up against prejudice.[37] Thus, our observed effects were found in situations where the victim or others are not present, and thus there is no diffusion of responsibility. If other bystanders are around, the motivated prejudice effect may be weaker given the heightened shared responsibility to confront would provide more external justification for not confronting, and thus generate less psychological discomfort for failing to confront, i.e., "others also could have confronted, but did not". If the victim is present, it can go both ways: the presence of the victim could heighten guilt and discomfort for not confronting, or actually less discomfort if bystanders would expect the victim to do something. Future studies ought to examine how these effects play out when the victim or other bystanders are present.

Beyond methodology, the intervention messages we tested in the second research has important boundary conditions. Namely, the beneficial effect of the induced loss mindset intervention was limited to those who were morally committed to non-prejudice. While we found an effective intervention only for some people, this is an important group of people to consider for promoting bystander confronting, because according to the motivated prejudice effect, those with more commitment to non-prejudice are especially likely to justify their inaction in face of prejudice through derogating the outgroup. They are also particularly relevant if taking into consideration cost-effectiveness of prosocial programs, given that non-prejudiced people are more likely to acknowledge and adopt messages communicating the importance of bystander intervention. Nevertheless, it remains an important direction for future research to conceive messages (most likely outside of the moral mindset domain) that would motivate confronting among those who are less committed to non-prejudice.

[37] This was different in only one study out of seven, in the moral mindset research, in one of the scenarios depicted in the vignette responsibility to intervene was shared with the other passengers on a bus.

Conclusion

In the current research, I focused on a route via which prejudiced sentiment in society can exponentially intensify over time. Specifically, I found that when bystanders choose not to confront a prejudiced perpetrator, albeit having an opportunity to, they themselves become more prejudiced (possibly in order to justify their inaction). Thereby creating a destructive cycle, where prejudice not confronted exponentially amplifies in a given social environment. Being aware of the various negative societal ramifications of prejudice expression, among them its direct effects on stigmatized victims, I aimed to create a psychological mindset intervention that motivates bystanders to speak up against prejudice. Knowing how intrapersonal costs of not confronting can lead to intergroup costs, I considered potential ways to use such anticipated intrapersonal costs to motivate moral behavior. Accordingly, I found that messages about prospective personal moral failure in regard to not intervening was potentially effective in promoting confrontation among non-prejudiced people. Given the growth of diverse societies, and occasional simultaneous rise in prejudiced discourse and atrocities, bystanders in the context of prejudice are becoming increasingly common, making the present research both timely and relevant.

References

Abad-Merino, S., Newheiser, A.K., Dovidio, J.F., Tabernero, C., & González, I. (2013). The dynamics of intergroup helping: The case of subtle bias against Latinos. *Cultural Diversity and Ethnic Minority Psychology, 19*(4), 445.

Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T. M., Rosenberg, M. J., & Tannenbaum, P. H. (Eds.). (1968). Theories of cognitive consistency: a sourcebook. Rand-McNally: Chicago.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm- glow giving. *The Economic Journal*, *100*(401), 464–477.

Aoki, J. (2015). *Seeking moral elevation vs. avoiding damnation: An examination of two moral motivational orientations* (Doctoral dissertation). Retrieved from Lehigh Preserve Database.

Apfelbaum, E. P., Norton, M. I., & Sommers, S. R. (2012). Racial color blindness: Emergence, practice, and implications. *Current directions in psychological science, 21*(3), 205-209.

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83,* 1423-1440.

Aronson, E., & Carlsmith, J. M. (1963). Effect of the severity of threat on the devaluation of forbidden behavior. *The Journal of Abnormal and Social Psychology*, *66*(6), 584.

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, *38*(2), 113-125.

Ashburn-Nardo, L. (2018). *What can allies do?* In A. J. Colella & E. B. King (Eds.), *Oxford library of psychology. The Oxford handbook of workplace discrimination* (p. 373–386). Oxford University Press.

Ashburn-Nardo, L., & Johnson, N. J. (2008). Implicit outgroup favoritism and intergroup judgment: The moderating role of stereotypic context. *Social Justice Research*, *21*(4), 490-508.

Ashburn-Nardo, L., & Karim, M. F. A. (2019). The CPR model: Decisions involved in confronting prejudiced responses. In *Confronting Prejudice and Discrimination* (pp. 29-47). Academic Press.

Ashburn-Nardo, L., Blanchar, J. C., Petersson, J., Morris, K. A., & Goodwin, S. A. (2014). Do you say something when it's your boss? The role of perpetrator power in prejudice confrontation. *Journal of Social Issues*, *70*(4), 615-636.

Ashburn-Nardo, L., Lindsey, A., Morris, K. A., & Goodwin, S. A. (2019). Who is responsible for confronting prejudice? The role of perceived and conferred authority. *Journal of Business and Psychology*, 1-13.

Ashburn-Nardo, L., Morris, K.A., & Goodwin, S.A. (2008). The confronting prejudiced responses (CPR) model: Applying CPR in organizations. *Academy of Management Learning & Education*, *7*(3), 332-342.

Ayres, M. M., Friedman, C. K., & Leaper, C. (2009). Individual and situational factors related to young women's likelihood of confronting sexism in their everyday lives. *Sex Roles*, *61*(7-8), 449-460.

Ball, T. C., & Branscombe, N. R. (2019). When do groups with a victimized past feel solidarity with other victimized groups?. In *Confronting prejudice and discrimination* (pp. 73-92). Academic Press.

Barany, Z. (2000). Politics and the Roma in state-socialist Eastern Europe. *Communist and Post-Communist Studies*, *33*, 421-437.

Barreto, M., & Ellemers, N. (2015). Detecting and experiencing prejudice: New answers to old questions. *Advances in experimental social psychology*, *52*, 139-219.

Batson, C. D. (2011). *Altruism in humans*. USA: Oxford University Press.

Batson, C. D., & Shaw, L. L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry*, *2*(2), 107–122.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183-200.

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

Bernát A, Juhász A, Krekó P and Molnár C (2013) 'The roots of radicalism and anti-Roma attitudes on the far right', in Kolosi T and Tóth I G (eds) Social Report 2012, TÁRKI: 355–376. http://www.tarki.hu/en/news/2013/items/20130305_bernat_juhasz _kreko_ molnar.pdf

Blanchard, F. A., Lilly, T., & Vaughn, L. A. (1991). Reducing the expression of racial prejudice. *Psychological Science*, *2*(2), 101-105.

Blanchard, F. A., Crandall, C. S., Brigham, J. C., & Vaughn, L. A. (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology*, *79*(6), 993.

Boysen, G. A. (2013). Confronting math stereotypes in the classroom: Its effect on female college students' sexism and perceptions of confronters. *Sex roles*, *69*(5-6), 297-307.

Brebels, L., De Cremer, D., Van Dijke, M., & Van Hiel, A. (2011). Fairness as social responsibility: A moral self-regulation account of procedural justice enactment. *British Journal of Management*, *22*, S47-S58.

Brinkman, B. G., Garcia, K., & Rickard, K. M. (2011). "What I wanted to do was…" discrepancies between college women's desired and reported responses to gender prejudice. *Sex Roles*, *65*(5-6), 344-355.

Burns, M. D., & Granz, E. L. (2021). Confronting Sexism: Promoting Confrontation Acceptance and Reducing Stereotyping through Stereotype Framing. *Sex Roles*, *84*(9), 503-521.

Burns, M. D., & Monteith, M. J. (2019). Confronting stereotypic biases: Does internal versus external motivational framing matter?. *Group Processes & Intergroup Relations*, *22*(7), 930-946.

Cameron, C.D., & Payne, B.K. (2012). The cost of callousness: Regulating compassion influences the moral self-concept. *Psychological Science*, *23*(3), 225-229.

Cadieux, J., & Chasteen, A. L. (2015). You gay, bro? Social costs faced by male confronters of antigay prejudice. *Psychology of Sexual Orientation and Gender Diversity*, *2*(4), 436.

Cascio, J., & Plant, E. A. (2015). Prospective moral licensing: Does anticipating doing good later allow you to be bad now?. *Journal of Experimental Social Psychology*, *56*, 110-116.

Chaney, K. E., & Sanchez, D. T. (2018). The endurance of interpersonal confrontations as a prejudice reduction strategy. *Personality and Social Psychology Bulletin*, *44*(3), 418-429.

Chaney, K. E., Young, D. M., & Sanchez, D. T. (2015). Confrontation's health outcomes and promotion of egalitarianism (C-HOPE) framework. *Translational Issues in Psychological Science*, *1*(4), 363.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817-869.

Chen, M. K., & Risen, J. L. (2010). How choice affects and reflects preferences: revisiting the free-choice paradigm. *Journal of Personality and Social Psychology*, *99*(4) ,573.

Cichocka, A., & Jost, J.T. (2014). Stripped of illusions? Exploring system justification processes in capitalist and post-Communist societies. *International Journal of Psychology*, *49*(1), 6-29.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*, 591-621.

Cihangir, S., Barreto, M., & Ellemers, N. (2014). Men as allies against sexism: The positive effects of a suggestion of sexism by male (vs. female) sources. *Sage Open*, *4*(2), 2158244014539168.

Cooper, J., & Fazio, R. H. (1984). A new look at dissonance theory. *Advances in Experimental Social Psychology, 17*, 229-266.

Craig, M. A., & Richeson, J. A. (2014). More diverse yet less tolerant? How the increasingly diverse racial landscape affects white Americans' racial attitudes. *Personality and Social Psychology Bulletin*, *40*(6), 750-761.

Crandall, C. S., Eshleman, A., & O'brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology, 82*(3), 359.

Crosby, J. R. (2006). Targeted social referencing and the perception of discrimination. *Dissertation Abstracts International: B. The Sciences and Engineering*, 67, 2874.

Crosby, J. R. (2015). The silent majority: Understanding and increasing majority group responses to discrimination. *Social and Personality Psychology Compass*, *9*(10), 539-550.

Crosby, J. R., & Wilson, J. (2015). Let's not, and say we would: Imagined and actual responses to witnessing homophobia. *Journal of Homosexuality*, *62*(7), 957-970.

Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially offensive behavior?. *Psychological Science*, *19*(3), 226-228.

Crowe, E., & Higgins, E. T. (1997). Regulatory focus and strategic inclinations: Promotion and prevention in decision-making. *Organizational Behavior and Human Decision Processes*, *69*(2), 117-132.

Csepeli, Gy.; Murányi, I., & Prazsák, G. (2011). *Új tekintélyelvűség Magyarországon.* Budapest: Apeiron.

Czopp, A. M. (2013). The passive activist: Negative consequences of failing to confront antienvironmental statements. *Ecopsychology*, *5*(1), 17-23.

Czopp, A. M. (2019). The consequences of confronting prejudice. In *Confronting Prejudice and Discrimination* (pp. 201-221). Academic Press.

Czopp, A. M., & Monteith, M. J. (2003). Confronting prejudice (literally): Reactions to confrontations of racial and gender bias. *Personality and Social Psychology Bulletin*, *29*(4), 532-544.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology*, *90*(5), 784.

Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited. *Personality and Social Psychology Bulletin*, *21*(11), 1139-1150.

Devine, P. G., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social psychology*, *60*(6), 817.

De Souza, L., & Schmader, T. (2021). The misjudgment of men: Does pluralistic ignorance inhibit allyship?. *Journal of Personality and Social Psychology*.

Dickter, C. L. (2012). Confronting hate: Heterosexuals' responses to anti-gay comments. *Journal of Homosexuality*, *59*(8), 1113-1130.

Dickter, C. L., & Newton, V. A. (2013). To confront or not to confront: Non-targets' evaluations of and responses to racist comments. *Journal of Applied Social Psychology*, *43*, E262-E275.

Dickter, C. L., Kittel, J. A., & Gyurovski, I. I. (2012). Perceptions of non-target confronters in response to racist and heterosexist remarks. *European Journal of Social Psychology*, *42*(1), 112-119.

Does, S., Derks, B., Ellemers, N., & Scheepers, D. (2012). At the heart of egalitarianism: how morality framing shapes cardiovascular challenge versus threat in Whites. *Social Psychological and Personality Science*, *3*(6), 747-753.

Dodd, E. H., Giuliano, T. A., Boutell, J. M., & Moran, B. E. (2001). Respected or rejected: Perceptions of women who confront sexist remarks. *Sex Roles*, *45*(7), 567-577.

Doosje, B., Branscombe, N. R., Spears, R., & Manstead, A. S. (1998). Guilty by association: When one's group has a negative history. *Journal of personality and social psychology*, *75*(4), 872.

Droogendyk, L., Wright, S. C., Lubensky, M., & Louis, W. R. (2016). Acting in solidarity: Cross-group contact between disadvantaged group members and advantaged group allies. *Journal of Social Issues*, *72*(2), 315-334.

Drury, B. J., & Kaiser, C. R. (2014). Allies against sexism: The role of men in confronting sexism. *Journal of social issues*, *70*(4), 637-652.

Dutton, D. G., & Lake, R. A. (1973). Threat of own prejudice and reverse discrimination in interracial situations. *Journal of Personality and Social Psychology*, *28*(1), 94.

Dutton, D. G., & Lennox, V. L. (1974). Effect of prior" token" compliance on subsequent interracial behavior. *Journal of Personality and Social Psychology*, *29*(1), 65.

Duval, S., & Wicklund, R. A. (1972). *A theory of objective self-awareness*. New York: Academic Press.

Effron, D. A., Cameron, J. S., & Monin, B. (2009). Endorsing Obama licenses favoring whites. *Journal of Experimental Social Psychology*, *45*(3), 590-593.

Eliezer, D., & Major, B. (2012). It’s not your fault: The social costs of claiming discrimination on behalf of someone else. *Group Processes & Intergroup Relations*, *15*(4), 487-502.

Fasoli, F., Paladino, M. P., Carnaghi, A., Jetten, J., Bastian, B., & Bain, P. G. (2016). Not “just words”: Exposure to homophobic epithets leads to dehumanizing and physical distancing from gay men. *European Journal of Social Psychology*, *46*(2), 237-248.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Fein, S., & Spencer, S.J. (1997). Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of Personality and Social Psychology*, *73*(1), 31.

Feischmidt, M., Szombati, K., & Szuhay, P. (2013) Collective criminalization of Roma in Central and Eastern Europe: Social causes, circumstances and consequences, In S. Body-Gendrot, M. Hough, K. Kerezsi, R. Levy, & S. Snacken S (Eds.), *The Routledge Handbook of European Criminology* (pp. 168-187). London, UK: Routledge.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fogelman, E. (n.d.). *The Rescuer Self.* Retrieved from: https://www.yadvashem.org/righteous/resources/the-rescuer-self.html

Freeman, D., Aquino, K., & McFerran, B. (2009). Overcoming beneficiary race as an impediment to charitable donations: Social dominance orientation, the experience of moral elevation, and donation behavior. *Personality and Social Psychology Bulletin, 35*(1), 72-84.

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio, & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism; prejudice, discrimination, and racism* (pp. 6189). San Diego, CA: Academic Press.

Garcia, D. M., Reser, A. H., Amo, R. B., Redersdorff, S., & Branscombe, N. R. (2005). Perceivers’ responses to in-group and out-group members who blame a negative outcome on discrimination. *Personality and Social Psychology Bulletin*, *31*(6), 769-780.

Gervais, S. J., & Hillard, A. L. (2014). Confronting sexism as persuasion: Effects of a confrontation's recipient, source, message, and context. *Journal of Social Issues*, *70*(4), 653-667.

Gervais, S. J., Hillard, A. L., & Vescio, T. K. (2010). Confronting sexism: The role of relationship orientation and gender. *Sex Roles*, *63*(7-8), 463-474.

Glasford, D. E., & Pratto, F. (2014). When extraordinary injustice leads to ordinary response: How perpetrator power and size of an injustice event affect bystander efficacy and collective action. *European Journal of Social Psychology*, *44*(6), 590-601.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*, 535- 549.

Gosling, P., Denizeau, M., & Oberlé, D. (2006). Denial of responsibility: a new mode of dissonance reduction. *Journal of Personality and Social Psychology*, *90*(5), 722.

Good, J. J., Moss-Racusin, C. A., & Sanchez, D. T. (2012). When do we confront? Perceptions of costs and benefits predict confronting discrimination on behalf of the self and others. *Psychology of Women Quarterly*, *36*(2), 210-226.

Good, J. J., Sanchez, D. T., & Moss-Racusin, C. A. (2018). A paternalistic duty to protect? Predicting men's decisions to confront sexism. *Psychology of Men & Masculinity*, *19*(1), 14.

Good, J. J., Woodzicka, J. A., Bourne, K. A., & Moss-Racusin, C. A. (2019). The decision to act: Factors that predict women's and men's decisions to confront sexism. In *Confronting Prejudice and Discrimination* (pp. 49-71). Academic Press.

Greenberg, J., & Pyszczynski, T. (1985). The effect of an overheard ethnic slur on evaluations of the target: How to spread a social disease. *Journal of Experimental Social Psychology*, *21*(1), 61-72.

Greenwald, A. G. (1975). On the inconclusiveness of "crucial" cognitive tests of dissonance versus self-perception theories. *Journal of Experimental Social Psychology*, *11*(5), 490-499.

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

Hayes, A. F. (2018). PROCESS (Version 3.1) [Computer software]. Retrieved from http://www.processmacro.org/download.html

Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods, 41*(3), 924-936.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist, 52*(12), 1280-1300.

Higgins, E. T., Roney, C. J., Crowe, E., & Hymes, C. (1994). Ideal versus ought predilections for approach and avoidance distinct self-regulatory systems. *Journal of Personality and Social Psychology, 66*(2), 276.

Hildebrand, L. K., Jusuf, C. C., & Monteith, M. J. (2020). Ally confrontations as identity-safety cues for marginalized individuals. *European Journal of Social Psychology*, *50*(6), 1318-1333.

Huff, C., & Tingley, D. (2015). “Who are these people?” Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, *2*(3), 1–12.

Huo, Y.J., Dovidio, J.F., Jiménez, T.R., & Schildkraut, D.J. (2018). Not just a national issue: Effect of state-level reception of immigrants and population changes on intergroup attitudes of Whites, Latinos, and Asians in the United States. *Journal of Social Issues, 74*(4),716-736.

Hyers, L. L. (2007). Resisting prejudice every day: Exploring women’s assertive responses to anti-Black racism, anti-Semitism, heterosexism, and sexism. *Sex Roles*, *56*(1), 1-12.

Johns, M., Schmader, T., & Lickel, B. (2005). Ashamed to be an American? The role of identification in predicting vicarious shame for anti-Arab prejudice after 9–11. *Self and Identity*, *4*(4), 331-348.

Karmali, F., Kawakami, K., & Page-Gould, E. (2017). He said what? Physiological and cognitive responses to imagining and witnessing outgroup racism. *Journal of Experimental Psychology: General*, *146*(8), 1073.

Kay, A.C., & Jost, J.T. (2003). Complementary justice: Effects of "poor but happy" and "poor but honest" stereotype exemplars on system justification and implicit activation of the justice motive. *Journal of Personality and Social Psychology, 85*(5), 823-837.

Kaiser, C. R., & Major, B. (2006). A social psychological perspective on perceiving and reporting discrimination. *Law & Social Inquiry*, *31*(4), 801-830.

Kaiser, C. R., & Miller, C. T. (2001). Stop complaining! The social costs of making attributions to discrimination. *Personality and Social Psychology Bulletin*, *27*, 254–263.

Kaiser, C. R., & Miller, C. T. (2003). Derogating the victim: The interpersonal consequences of blaming events on discrimination. *Group Processes & Intergroup Relations*, *6*(3), 227-237.

Kaiser, C. R., Hagiwara, N., Malahy, L. W., & Wilkins, C. L. (2009). Group identification moderates attitudes toward ingroup members who confront discrimination. *Journal of Experimental Social Psychology*, *45*(4), 770-777.

Kant, I. (1785). *Groundwork of the metaphysic of morals*.

Karmali, F., Kawakami, K., & Page-Gould, E. (2017). He said what? Physiological and cognitive responses to imagining and witnessing outgroup racism. *Journal of Experimental Psychology: General*, *146*(8), 1073.

Katz, J., Federici, D., & Brown, D. (2021). Effects of Humor and Bystander Gender on Responses to Antigay Harassment. *Journal of Homosexuality*, 1-20.

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J.F. (2009). Mispredicting affective and behavioral responses to racism. *Science*, *323*(5911), 276-278.

Kawakami, K., Karmali, F., & Vaccarino, E. (2019). Confronting intergroup bias: Predicted and actual responses to racism and sexism. In *Confronting Prejudice and Discrimination* (pp. 3-28). Academic Press.

Kemény, I., Lengyel, I., & Janky, B. G. (2004). *A magyarországi cigányság: 1971-2003,* Budapest: Gondolat.

Kende, A., Hadarics, M., & Lášticová, B. (2017). Anti-Roma attitudes as expressions of dominant social norms in Eastern Europe. *International Journal of Intercultural Relations*, *60*, 12-27.

Kende, A., Nyúl, B., Hadarics, M., Wessenauer, V., & Hunyadi, B. (2018) *Antigypsyism and Antisemitism in Hungary. Summary of the final report*. Retrieved from: https://politicalcapital.hu/pc-admin/source/documents/EVZ_Antigypsyism%20Antisemitism_final%20report_%20summary_180228.pdf

Kerr, N. L., & Kaufman-Gilliland, C. M. (1997). ".. and besides, I probably couldn't have made a difference anyway": Justification of Social Dilemma Defection via Perceived Self-Inefficacy. *Journal of Experimental Social Psychology, 33*(3), 211-230.

Kovács, A. (2002). *Zsidók és zsidóság a mai Magyarországon. Egy szociológiai kutatás eredményei*. Budapest: Szombat.

Kovács, A. (2010). *Stranger at hand: Antisemitic prejudices in post-communist Hungary.* Boston & Leiden: Brill.

Kovács, A. (2014) Zsidóellenes előítéletesség és az antiszemitizmus dinamikája a mai Magyarországon, Kolosi T. és Tóth I.Gy. (szerk). *Társadalmi Riport*, *12*, 486-508.

Kovács, A., & Barna, I. (2018). *Zsidók és zsidóság Magyarországon 2017: Egy szociológiai kutatás eredményei*. Budapest: Szombat.

Kroeper, K. M., Sanchez, D. T., & Himmelstein, M. S. (2014). Heterosexual men's confrontation of sexual prejudice: The role of precarious manhood. *Sex Roles*, *70*(1-2), 1-13.

Krolikowski, A. M., Rinella, M., & Ratcliff, J. J. (2016). The influence of the expression of subtle and blatant sexual prejudice on personal prejudice and identification with the expresser. *Journal of homosexuality, 63*(2), 228-249.

Kteily, N., & Bruneau, E. (2017) Backlash: The politics and real-world consequences of minority group dehumanization. *Personality and Social Psychology Bulletin, 43*(1), 87-104.

Kunst, J. R., Sam, D. L., & Ulleberg, P. (2013). Perceived islamophobia: Scale development and validation. *International Journal of Intercultural Relations, 37*(2), 225-237.

Kutlaca, M., Becker, J., & Radke, H. (2020). A hero for the outgroup, a black sheep for the ingroup: Societal perceptions of those who confront discrimination. *Journal of Experimental Social Psychology, 88*, 103832.

Ladányi, J. (2001). The Hungarian neoliberal state, ethnic classification, and the creation of a Roma underclass. In R. J. Emigh & I Szelényi (Eds.) *Poverty, ethnicity, and gender in Eastern Europe during the market transition*, (pp. 67-82). Westport, CT: Greenwood Publishing Group.

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

Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review of Psychology, 58*, 317-344.

Lee RT, Perez AD, Boykin CM, Mendoza-Denton R (2019) On the prevalence of racial discrimination in the United States. *PLoS ONE, 14*(1).

Lee, S. A., Reid, C. A., Short, S. D., Gibbons, J. A., Yeh, R., & Campbell, M. L. (2013). Fear of Muslims: Psychometric evaluation of the Islamophobia Scale. *Psychology of Religion and Spirituality, 5*(3), 157.

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science, 22*(12), 1472-1477.

Leippe, M. R., & Eisenstadt, D. (1994). Generalization of dissonance reduction: Decreasing prejudice through induced compliance. *Journal of Personality and Social Psychology, 67*(3), 395.

Leippe, M. R., & Eisenstadt, D. (1999). A self-accountability model of dissonance: Multiples modes on continuum of elaboration. In E. Harmon-Jones & J. Mills (Eds.), *Cognitive dissonance: Progress on a pivotal theory in social psychology* (pp. 201–232). Washington, DC: American Psychological Association.

Lerner, M. J., & Simmons, C.H. (1966). Observer's reaction to the" innocent victim": Compassion or rejection?. *Journal of Personality and Social Psychology*, *4*(2), 203.

Lewis, T., & Yoshimura, S. M. (2017). Politeness Strategies in Confrontations of Prejudice. *Atlantic Journal of Communication*, *25*(1), 1-16.

Lickel, B., Schmader, T., Curtis, M., Scarnier, M., & Ames, D. R. (2005). Vicarious shame and guilt. *Group Processes & Intergroup Relations*, *8*(2), 145-157.

Lindsey, A., King, E., Cheung, H., Hebl, M., Lynch, S., & Mancini, V. (2015). When do women respond against discrimination? Exploring factors of subtlety, form, and focus. *Journal of Applied Social Psychology*, *45*(12), 649-661.

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 1–10.

Ljujic, V., Vedder, P., Dekker, H., & van Geel, M. (2012). Romaphobia: A unique phenomenon?. *Romani Studies*, *22*, 141-152.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122-1135.

Mallett, R. K., & Melchiori, K. J. (2014). Goal preference shapes confrontations of sexism. *Personality and Social Psychology Bulletin*, *40*(5), 646-656.

Mallett, R. K., & Melchiori, K. J. (2019). Goals drive responses to perceived discrimination. In *Confronting Prejudice and Discrimination* (pp. 95-119). Academic Press.

Mallet, R. K., & Monteith, M. J. (Eds.) (2019). *Confronting prejudice and discrimination: The science of changing minds and behaviors*. London, UK: Academic Press.

Mallett, R. K., & Wagner, D. E. (2011). The unexpectedly positive consequences of confronting sexism. *Journal of Experimental Social Psychology*, *47*(1), 215-220.

Mallett, R. K., Ford, T. E., & Woodzicka, J. A. (2016). What did he mean by that? Humor decreases attributions of sexism and confrontation of sexist jokes. *Sex Roles*, *75*(5), 272-284.

Mallett, R. K., Ford, T. E., & Woodzicka, J. A. (2019). Ignoring sexism increases women's tolerance of sexual harassment. *Self and Identity*, 1-17.

McGrath, A. (2017). Dealing with dissonance: A review of cognitive dissonance reduction. *Social and Personality Psychology Compass*, *11*(12).

Miller, W. I. (2000). *The mystery of courage*. Cambridge, MS: Harvard University Press.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*(1), 33.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology*, *65*(3), 469.

Monteith, M. J., Burns, M. D., & Hildebrand, L. K. (2019). Navigating successful confrontations: What should I say and how should I say it?. In *Confronting prejudice and discrimination* (pp. 225-248). Academic Press.

Monteith, M. J., Deneen, N. E., & Tooman, G. D. (1996). The effect of social norm activation on the expression of opinions concerning gay men and Blacks. *Basic and Applied Social Psychology*, *18*(3), 267-288.

Munder, A. K., Becker, J. C., & Christ, O. (2020). Standing up for whom? Targets' different goals in the confrontation of discrimination. *European Journal of Social Psychology*, *50*(7), 1443-1462.

Myers, D. G. (1975). Discussion-induced attitude polarization. *Human Relations*, *28*(8), 699-714.

Nietzsche, F. (1878). *Human, all too human*.

Nguyen, H. H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93*(6), 1314.

Norton, M. I., Sommers, S. R., Apfelbaum, E. P., Pura, N., & Ariely, D. (2006). Color blindness and interracial interaction: Playing the political correctness game. *Psychological Science, 17*(11), 949-953.

O'Brien, L. T., Crandall, C. S., Horstman-Reser, A., Warner, R., Alsbrooks, A., & Blodorn, A. (2010). But I'm no bigot: How prejudiced White Americans maintain unprejudiced self-images. *Journal of Applied Social Psychology*, *40*(4), 917-946.

Oliver, M. B. (2003). African American men as" criminal and dangerous": Implications of media portrayals of crime on the" criminalization" of African American men. *Journal of African American Studies*, 3-18.

Oswald, D. L. (2005). Understanding Anti-Arab Reactions Post-9/11: The Role of Threats, Social Categories, and Personal Ideologies. *Journal of Applied Social Psychology*, *35*, 1775–1799.

Osswald, S., Greitemeyer, T., Fischer, P., & Frey, D. (2010). Moral prototypes and moral behavior: Specific effects on emotional precursors of moral behavior and on moral behavior by the activation of moral prototypes. *European Journal of Social Psychology*, *40*(6), 1078–1094.

Pascoe, E. A., & Smart Richman, L. (2009). Perceived discrimination and health: a meta-analytic review. *Psychological bulletin*, *135*(4), 531.

Paradies, Y., Ben, J., Denson, N., Elias, A., Priest, N., Pieterse, A., ... & Gee, G. (2015). Racism as a determinant of health: a systematic review and meta-analysis. *PloS one, 10*(9), e0138511.

Parker, L. R., Monteith, M. J., Moss-Racusin, C. A., & Van Camp, A. R. (2018). Promoting concern about gender bias with evidence-based confrontation. *Journal of Experimental Social Psychology*, *74*, 8-23.

Pew Research Center (2021). Internet/Broadband Fact Sheet. Retrieved from: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

Plant, E. A., & Butz, D. A. (2006). The causes and consequences of an avoidance-focus for interracial interactions. *Personality and Social Psychology Bulletin*, *32*(6), 833-846.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*(3), 811.

Plant, E. A., & Devine, P. G. (2001). Responses to other-imposed pro-Black pressure: Acceptance or backlash?. *Journal of Experimental Social Psychology*, *37*(6), 486-501.

Plant, E. A., & Devine, P. G. (2009). The active control of prejudice: unpacking the intentions guiding control efforts. *Journal of Personality and Social Psychology, 96*(3), 640.

Plous, S. (2000). Responding to overt displays of prejudice: A role-playing exercise. *Teaching of Psychology*, *27*(3), 198-200.

Pogány, I. (2006). Minority rights and the Roma of Central and Eastern Europe. *Human Rights Law Review*, *6*, 1-25. doi: 10.1093/hrlr/ngi034

Poteat, V. P., & Vecho, O. (2016). Who intervenes against homophobic behavior? Attributes that distinguish active bystanders. *Journal of school psychology*, *54*, 17-28.

Powell, A. A., Branscombe, N. R., & Schmitt, M. T. (2005). Inequality as ingroup privilege or outgroup disadvantage: The impact of group focus on collective guilt and interracial attitudes. *Personality and Social Psychology Bulletin, 31*(4), 508-521.

Pratto, F., Cidam, A., Stewart, A. L., Zeineddine, F. B., Aranda, M., Aiello, A., ... & Eicher, V. (2013). Social dominance in context and in individuals: Contextual moderation of robust effects of social dominance orientation in 15 languages and 20 countries. *Social Psychological and Personality Science*, *4*(5), 587-599.

Quinn, K. A., & Olson, J. M. (2011). Regulatory framing and collective action: The interplay of individual self-regulation and group behavior. *Journal of Applied Social Psychology*, *41*(10), 2457-2478.

Radke, H. R., Kutlaca, M., Siem, B., Wright, S. C., & Becker, J. C. (2020). Beyond allyship: Motivations for advantaged group members to engage in action for disadvantaged groups. *Personality and Social Psychology Review*, *24*(4), 291-315.

Rasinski, H. M., & Czopp, A. M. (2010). The effect of target status on witnesses' reactions to confrontations of bias. *Basic and Applied Social Psychology*, *32*(1), 8-16.

Rasinski, H. M., Geers, A. L., & Czopp, A. M. (2013). “I guess what he said wasn’t that bad” dissonance in nonconfronting targets of prejudice. *Personality and Social Psychology Bulletin, 39*(7), 856-869.

Rattan, A. (2019). How lay theories (or mindsets) shape the confrontation of prejudice. In *Confronting Prejudice and Discrimination* (pp. 121-140). Academic Press.

Rattan, A., & Dweck, C. S. (2010). Who confronts prejudice? The role of implicit theories in the motivation to confront prejudice. *Psychological Science*, *21*(7), 952-959.

Reimer, N. K., Becker, J. C., Benz, A., Christ, O., Dhont, K., Klocke, U., ... & Hewstone, M. (2017). Intergroup contact and social change: Implications of negative and positive contact for collective action in advantaged and disadvantaged groups. *Personality and Social Psychology Bulletin*, *43*(1), 121-136.

Saguy, T., & Szekeres, H. (2018). Changing Minds via Collective Action: Exposure to the 2017 Women's March Predicts Over-time Decrease in (Some) Men's Gender System Justification. *Group Processes & Intergroup Relations, 21(5*), 678-689.

Sassenberg, K., & Hansen, N. (2007). The impact of regulatory focus on affective responses to social discrimination. *European Journal of Social Psychology*, *37*(3), 421-444.

Satpute, A.B., Kragel, P.A., Barrett, L.F., Wager, T.D., & Bianciardi, M. (2019). Deconstructing arousal into wakeful, autonomic and affective varieties. *Neuroscience Letters, 693,* 19-28.

Schaumberg, R. L., & Wiltermuth, S. S. (2014). Desire for a positive moral self-regard exacerbates escalation of commitment to initiatives with prosocial aims. *Organizational Behavior and Human Decision Processes*, *123*(2), 110-123.

Schmader, T., Croft, A., Scarnier, M., Lickel, B., & Mendes, W. B. (2012). Implicit and explicit emotional reactions to witnessing prejudice. *Group Processes & Intergroup Relations*, *15*(3), 379-392.

Schmitt, M. T., Branscombe, N. R., Postmes, T., & Garcia, A. (2014). The consequences of perceived discrimination for psychological well-being: a meta-analytic review. *Psychological Bulletin*, *140*(4), 921.

Scholer, A. A., Zou, X., Fujita, K., Stroessner, S. J., & Higgins, E. T. (2010). When risk seeking becomes a motivational necessity. *Journal of Personality and Social Psychology*, *99*(2), 215.

Schultz, J. R., & Maddox, K. B. (2013). Shooting the messenger to spite the message? Exploring reactions to claims of racial bias. *Personality and Social Psychology Bulletin*, *39*(3), 346-358.

Sechrist, G. B. (2010). Making attributions to and plans to confront gender discrimination: The role of optimism. *Journal of Applied Social Psychology*, *40*(7), 1678-1707.

Sedikides, C. (1993). Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of Personality and Social Psychology*, *65*(2), 317.

Sedikides, C., & Hepper, E. G. (2009). Self-improvement. *Social and Personality Psychology Compass*, *3*(6), 899-917.

Shah, J., & Higgins, E. T. (1997). Expectancy× value effects: Regulatory focus as determinant of magnitude and direction. *Journal of Personality and Social Psychology*, *73*(3), 447.

Shelton, J. N., & Stewart, R. E. (2004). Confronting perpetrators of prejudice: The inhibitory effects of social cost. *Psychology of Women Quarterly*, 215–223.

Shelton, J. N., Richeson, J. A., Salvatore, J., & Hill, D. M. (2006). Silence is not golden: Intrapersonal consequences of not confronting prejudice. In S. Levin & C. Van Laar (Eds.), *Social stigma and group inequality: Social psychological perspectives*. Mahwah, NJ: Erlbaum.

Shelton, J. N., West, T. V., & Trail, T. E. (2010). Concerns about appearing prejudiced: Implications for anxiety during daily interracial interactions. *Group Processes & Intergroup Relations, 13*(3), 329-344.

Sik, E. & Simonovits, B. (2012). *A diszkrimináció mérése. e-tankönyv*. Budapest, ELTE. Retrieved from: https://www.tarki.hu/hu/about/staff/sb/Diszkriminacio_merese.pdf

Simon, L., & Greenberg, J. (1996). Further progress in understanding the effects of derogatory ethnic labels: The role of preexisting attitudes toward the targeted group. *Personality and Social Psychology Bulletin*, *22*(12), 1195-1204.

Simon, L., Greenberg, J., & Brehm, J. (1995). Trivialization: the forgotten mode of dissonance reduction. *Journal of Personality and Social Psychology*, *68*(2), 247.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, *4*(4), 267-281.

Skitka, L. J. (2012). Moral convictions and moral courage: Common denominators of good and evil. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 349-365). Washington, D.C: American Psychological Association.

Skitka, L. J. (2014). The psychological foundations of moral conviction. *Advances in Experimental Moral Psychology, 148*.

Skitka, L. J., & Morgan, G. S. (2014). The social and political implications of moral conviction. *Political Psychology*, *35*, 95-110.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more?. *Journal of Personality and Social Psychology*, *88*(6), 895.

Sommers, S. R., & Norton, M. I. (2006). Lay theories about White racists: What constitutes racism (and what doesn't). *Group Processes & Intergroup Relations*, *9*(1), 117-138.

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior*, *44*(2), 136-146.

Stangor, C., Swim, J. K., Van Allen, K. L., & Sechrist, G. B. (2002). Reporting discrimination in public and private contexts. *Journal of Personality and Social Psychology*, *82*(1), 69.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797.

Stone, J., & Cooper, J. (2001). A self-standards model of cognitive dissonance. *Journal of Experimental Social Psychology*, *37*(3), 228-243.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: implications for clinical practice. *American psychologist*, *62*(4), 271.

Swim, J. K., & Hyers, L. L. (1999). Excuse me—What did you just say?!: Women's public and private responses to sexist remarks. *Journal of Experimental Social Psychology*, *35*(1), 68-88.

Swim, J. K., & Thomas, M. A. (2006). Responding to everyday discrimination: A synthesis of research on goal-directed, self-regulatory coping behaviors. *Stigma and group inequality: Social Psychological Perspectives*, 105-126.

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

Szekeres, H. (2020). Kedvelni vagy tisztelni? Az előítéletek csökkentése a melegszívűség és kompetencia dimenzióiban. *Alkalmazott Pszichológia*, *20*(4), 123-157.

Szekeres, H., Shuman, E., & Saguy, T. (2020). Views of sexual assault following# MeToo: The role of gender and individual differences. *Personality and Individual Differences, 166*, 110203.

Torres, L., Reveles, A. K., Mata-Greve, F., Schwartz, S., & Domenech Rodriguez, M. M. (2020). Reactions to Witnessing Ethnic Microaggressions: An Experimental Study. *Journal of Social and Clinical Psychology, 39*(2), 141-164.

Triana, M. D. C., Jayasinghe, M., & Pieper, J. R. (2015). Perceived workplace racial discrimination and its correlates: A meta-analysis. *Journal of Organizational Behavior*, *36*(4), 491-513.

Tykocinski, O. E., Pittman, T. S., & Tuttle, E. E. (1995). Inaction inertia: Foregoing future benefits as a result of an initial failure to act. *Journal of Personality and Social Psychology*, *68*(5), 793.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039-1061.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297-323.

Twenge, J. M., & Crocker, J. (2002). Race and self-esteem: Meta-analyses comparing whites, blacks, Hispanics, Asians, and American Indians and comment on Gray-Little and Hafdahl (2000). *Psychological Bulletin, 128*, 371–408.

Valentino, N.A., Brader, T. & Jardina, A.E. (2013). Immigration opposition among US Whites: General ethnocentrism or media priming of attitudes about Latinos?. *Political Psychology, 34*(2),149-166.

Van Zomeren, M. (2013). Four core social-psychological motivations to undertake collective action. *Social and Personality Psychology Compass, 7*(6), 378-388.

Van Zomeren, M., & Iyer, A. (2009). Introduction to the social and psychological dynamics of collective action. *Journal of Social Issues, 65*(40), 645-660.

Van Zomeren, M., Kutlaca, M., & Turner-Zwinkels, F. (2018). Integrating who “we” are with what “we”(will not) stand for: A further extension of the Social Identity Model of Collective Action. *European Review of Social Psychology*, *29*(1), 122-160.

Van Zomeren, M., Postmes, T., & Spears, R. (2008). Toward an integrative social identity model of collective action: a quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin, 134*(4), 504.

Voils, C. I., Ashburn-Nardo, L., & Monteith, M.J. (2002). Evidence of prejudice-related conflict and associated affect beyond the college setting. *Group Processes & Intergroup Relations*, *5*(1), 19-33.

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: race, social fit, and achievement. *Journal of personality and social psychology*, *92*(1), 82.

Wang, K., & Dovidio, J. F. (2017). Perceiving and confronting sexism: The causal role of gender identity salience. *Psychology of Women Quarterly, 41*(1), 65-76.

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the

PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063.

Wellman, J. A., Czopp, A. M., & Geers, A. L. (2009). The egalitarian optimist and the confrontation of prejudice. *The Journal of Positive Psychology*, *4*(5), 389-395.

Williams, D. R., Lawrence, J. A., & Davis, B. A. (2019). Racism and health: evidence and needed research. *Annual review of public health*, *40*, 105-125.

Wills, T. A. (1981). Downward comparison principles in social psychology. *Psychological Bulletin*, *90*(2), 245.

Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science, 14*(3), 131-134.

Winterich, K. P., Aquino, K., Mittal, V., & Swartz, R. (2013a). When moral identity symbolization motivates prosocial behavior: The role of recognition and moral identity internalization. *Journal of Applied Psychology*, *98*(5), 759.

Winterich, K. P., Mittal, V., & Aquino, K. (2013). When does recognition increase charitable behavior? Toward a moral identity-based model. *Journal of Marketing, 77*(3), 121-134.

Woodzicka, J. A., & LaFrance, M. (2001). Real versus imagined gender harassment. *Journal of Social Issues*, *57*(1), 15-30.

Wright, S. C. (2009). The next generation of collective action research. *Journal of Social Issues, 65*, 859–879.

Zaal, M. P., Laar, C. V., Ståhl, T., Ellemers, N., & Derks, B. (2011). By any means necessary: The effects of regulatory focus and moral conviction on hostile and benevolent forms of collective action. *British Journal of Social Psychology, 50*(4), 670-689.

Zaal, M. P., Van Laar, C., Ståhl, T., Ellemers, N., & Derks, B. (2012). Social change as an important goal or likely outcome: How regulatory focus affects commitment to collective action. *British Journal of Social Psychology*, *51*(1), 93-110.

Zitek, E. M., & Hebl, M. R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology, 43*(6), 867-876.

Zou, L. X., & Dickter, C. L. (2013). Perceptions of racial confrontation: The role of color blindness and comment ambiguity. *Cultural Diversity and Ethnic Minority Psychology*, *19*(1), 92.

Zuwerink, J.Z., Devine, P.G., Monteith, M. J., & Cook, D.A. (1996). Prejudice toward Blacks: With and without compunction?. *Basic and Applied Social Psychology*, *18*(2), 131-150.

Appendices

Appendix A: Materials of Chapter 2

Demographic questions

Study 1 (HU study):

Age: _________
Gender:
- o Male
- o Female
- o Other: ____
- o I don’t wish to answer

Your education level:
- o primary education or lower
- o secondary education
- o ongoing university education
- o completed university education

What is your denomination?
- o Catholic
- o Reformed
- o Evangelist
- o Jewish
- o Muslim
- o Other
- o None
- o I don’t wish to answer

Please indicate which group you most identify with?
- o Roma
- o Schwab [German origin]
- o Serbian
- o Slovakian
- o Jewish
- o Hungarian outside of the border
- o Majority “average” Hungarian
- o None of the above
- o Other: ___________

People often describe their political orientation along the dimensions of left and right, liberal and conservative. Please indicate where you would place yourself along these dimensions.

Left-wing o o o o o o o Right-wing [bipolar scale]
Liberal o o o o o o o Conservative [bipolar scale]

How do you judge your economic status to be like?
- o I live much worse than the average
- o I live worse than the average

- I live a bit worse than the average
- Average
- I live a bit better than the average
- I live better than the average
- I live much better than the average

Study 2 and 3 (US studies):

Age: _________

Gender:

- Male
- Female
- Other: ____
- I do not wish to answer

Your education level [analyzed as continuous variable by excluding ‘other’]:

- less than high school
- high school diploma
- bachelor’s degree
- master’s degree
- PhD
- other

What is your race or ethnicity?

- White/Caucasian/European American
- Black/African American/Afro-Caribbean
- Latino/Hispanic
- Native American
- Asian
- Arab
- Biracial/Mixed: ______
- Other

What is your religion?

- No religious affiliation
- Christian
- Muslim
- Jewish
- Hindu
- Buddhist
- Other

Compared to other people in your society, what is your economic situation? (check one)

- destitute
- poor
- so-so
- good
- better than most
- wealthy

[Study 3] Compared to other people in your society, what is your socio-economic situation?

- o Much worse than average (1)
- o Worse than average (2)
- o Average (3)
- o Better than average (4)
- o Much better than average (5)

What is your socio-political orientation?

Conservative Liberal
0 --------------------------------------10 [slider]

What is your political affiliation?

- o Democrat
- o Republican
- o Neither
- o Don’t want to answer

Who did you vote for in the recent [“last” in Study 3] presidential elections?

- o Hillary Clinton
- o Donald Trump
- o Other candidate: ____________
- o I did not vote
- o I do not wish to answer

Socio-political–intergroup attitudes

Political ideology and affiliation questions are included in demographic.

Political antisemitism scale (Study 1)
Shortened and adapted from Kovács, 2014
To what extent do you agree or disagree with the following statements:
(On a 7-point Likert scale from 1 = strongly disagree to 7 = strongly agree)

1. Jewish influence is too great in Hungary today.
2. It would be best if the Jews emigrated from the country.
3. The number of Jews in certain fields of employment should be restricted.
4. The Jews are more prone to using unclean means to achieve their goals than others.
5. Intellectuals of Jewish origin keep media and culture under their influence.
6. There is a secret Jewish conspiracy that determines political and economic processes.

System Justification scale (Study 2-3)
Shortened and adapted from Kay & Jost, 2003
To what extent do you agree with the followings:
(On a 9-point Likert scale from 1 = strongly disagree to 9 = strongly agree)

1. In general, I find society to be fair

2. American society needs to be radically restructured [reverse-scored]
3. Most policies serve the greater good
4. Everyone has a fair shot at wealth and happiness
5. Society is set up so that people usually get what they deserve.

SDO scale (Study 2-3)

SDO scale (SSDO) taken from Pratto et al., 2013

There are many kinds of groups in the world: men and women, ethnic and religious group, nationalities, political factions. How much do you support or oppose the ideas about groups in general?

(On a 10-point Likert scale from 1 = *extremely oppose* to 10 = *extremely favor*)

1. In setting priorities, we must consider all groups. (reversed-scored)
2. We should not push for group equality.
3. Group equality should be our ideal. (reversed-scored)
4. Superior groups should dominate inferior group.

IMS scale (Study 2-3)

Shortened and adapted from Plant & Devine, 1998

To what extent do you agree or disagree with the following statements:
(On a 9-point Likert scale from 1 = strongly disagree to 9 = strongly agree)

1. I am personally motivated by my beliefs to be nonprejudiced toward Muslims/Latinos.
2. Because of my personal values, I believe that using stereotypes about Muslims/Latinos is wrong.
3. Being nonprejudiced toward Muslims/Latinos is important to my self-concept.
4. I attempt to act in nonprejudiced ways toward Muslims/Latinos because it is personally important to me.

(reversed-score item from original scale was not included: “According to my personal values, using stereotypes about Black people is OK.”)

Modern Racism Scale (MRS; Study 3)

(On 5-point scale, 1=strongly disagree, 2=somewhat disagree, 3=neither agree nor disagree, 4=somewhat agree, 5= strongly agree; α = .93):

1. There are too many foreign students of Hispanic descent being allowed to attend university in the U.S.;
2. The U.S. should open its doors to more Latino immigration from the poorer countries. (reversed-scored);
3. It’s good to live in a country where there are so many Latinos. (reversed-scored);
4. Intermarriage between Latinos and Whites is a good thing for the U.S. (reversed-scored);

5. It is not fair that so many scholarships and awards are awarded to Latino students.;
6. It is too easy for Latinos to illegally arrive in the U.S.;
7. Many Latinos do not bother to learn proper English.;
8. Discrimination against Latinos is no longer a problem in the U.S.;
9. White Americans do not get treated very well in places dominated by Latinos.

Game instructions

The Trust Game description and instructions (Study 1-2, Study 5)

You are going to observe others playing a game called the **Trust Game**, then you will have a chance to play it also (you can win money in this game).

We are testing:
- How does someone observing us influence our trusting behavior?
- How does the gender of the observer influence trusting behavior?
- Are women or men more trustworthy?
- Are people more trustworthy towards women or men?

You'll be assigned to observe one player and his/her rounds.
Note that this observed player will know you are watching and has an opportunity to send you private messages after each round (we allowed this option so the observer feels more real to the player). You will have a chance to message him/her back. **Make sure that whatever you say, you do not reveal the purpose of the study to the player you are corresponding with.**

Pay attention to the game - 1, the player might message you and 2, it is beneficial for you to get familiar with the game before you are playing it yourself.

<u>How do you play the Trust Game?</u>

- Two people are in the game. An amount is given to the first player (for example, $2).
- The first player can decide between three options, to give "ALL" ($2), "HALF" ($1), or "NONE" ($0 = keep all the $2) of this amount to the other player. Let's say s/he decides to give HALF ($1).
- This amount is tripled automatically and given to the opponent ($3).
- The opponent can decide between two options, to "SHARE" the amount with the first player (half it, $1.50) or give back "NONE" (keep all the $3). Let's say the opponent decides to SHARE ($1.50)
- The game ends.
- The first player won $1 (since s/he kept half of the amount to him/herself in the beginning) plus $1.50 (which the opponent gave back in the end), ending up in $2.50.

- The opponent won $1.50.

Note, players earn the most money if the first player trusts the opponent and gives "ALL", and the opponent is fair and chooses to "SHARE" (giving half of the earned money back). In this case, following the previous example, they would both end up with $3.

Meaning, you can earn the most if you give it ALL and the other player is trustworthy and chooses SHARE.

The Share Game description and instructions (Study 3)

You are going to observe others playing a game called the Share Game.
Then you will play shortly and you can earn bonus money.

<u>How do you play the Share Game?</u>

- There is a "giver" and "receiver" at each round.
- The "giver" is provided with money, and he/she can decide to either share (half of) it with the opponent, the "receiver", or give none and keep it all to herself.
- A player can be a "giver" in one round, and a "receiver" in another, so the roles change around among players, and even the same opponents might change roles.
- So the "giver" may share out of sense of fairness, but also because in another round, she may become a "receiver" to someone she either shared with or did not share.

In the Share Game, on average, players earn more money if they keep sharing. As a rule of thumb, players should share.

We are testing the influence of third-party presence and gender. How does someone observing us influence sharing behavior? How does the gender of the observer influence sharing behavior? Are women or men more sharing? Are people more sharing toward women or men?

You'll be assigned to **observe 1 player** and his/her rounds. Once you observed the player, you will play 1 round with her, where she will be the "giver" and you will be the "receiver".

Important:

- The player you are observing has an opportunity to send you private messages.
- We allowed this option so the observer feels more real to the player, *but* to limit conversation, you can't initiate messaging, only reply.

- There is no messaging between players, so you won't receive messages or be able to reply *while* you play with the observed player.
- **Do not reveal the purpose of the study to the observed player!**

Please keep your response to a minimum!

PANAS Negative affect scale and Psychological discomfort scale (Study 3)

The extent you are feeling this way at the moment…
(on a 5-point scale, where 1=very slightly or not at all, 2=a little, 3=moderately, 4=quite a bit, 5=extremely)
Afraid*
Ashamed*
Bothered**
Distressed*
Guilty*
Hostile*
Irritable*
Jittery*
Nervous*
Scared*
Upset*
Uncomfortable**
Uneasy**

** PANAS negative subscale, adapted from Watson, Clark, & Tellegen, 1988*
*** Psychological discomfort scale adapted from Elliot & Devine, 1994*

Disclosure lead up to trivialization measure

Study 1

Intergroup conditions:
According to one of the participants a player seemed prejudiced. Did you observe or experience ethnic/religion-based prejudice or discrimination from any players?
Interpersonal condition:
A participant told us that one of the players they observed, offered zero money (nothing) to their partner player (who had a stereotypically Jewish name), and then privately messaged the participant saying: “well I won’t trust these rothschilds with money”. Have you witnessed such and similar treatment?
[Translated]

Study 2

Intergroup condition:

We asked if you had witnessed anything insulting because a respondent informed us about a player commenting about another player's ethnicity. Did you encounter such behavior?

Interpersonal condition:

We asked if you had witnessed anything insulting because another respondent who was observing the game informed us that there was a player who offered no money to his opponent who had an Arab/Muslim name and then messaged the respondent saying that "You can't trust those damn muslims". Did you witness such behavior?

Study 3

Interpersonal condition:

Another respondent informed us that there was a player who offered no money to his opponent who had a Latino name, and then messaged the respondent saying that "yeah like if you could only trust latinos not stealinh our jobs". Did you encounter such behavior?

Debriefing script

Study 1

Dear participant,

In our research, we were interested to learn, how do people behave in uncomfortable intergroup situations and what are the consequences of these behaviors. We kindly ask you to please not to share for now the purpose of this study with your mates at the university, because they may take part in this experiment in the future as well.

If you have any questions or comments, please write us an email at: socialpsychology.research@yahoo.com-ra

Study 2-3

Thank you again for participating in this psychological study!

With the payment you are receiving a 50 cents bonus as promised.

We were interested in learning how people react emotionally, behaviorally and cognitively to uncomfortable intergroup situations, such as to witnessing racism. To this end, we created a situation that is not real - the game you have seen was programmed. Some of you witnessed a player being inappropriate. We apologize for deceiving you this way, but we believe it is important to learn about the considerations people have when deciding to confront or not to confront in such social situations. We apologize if we made you feel uncomfortable in any way.

Any data collected in this study will be reviewed **anonymously** for research purposes only!

We trust you not to share information about this study with other mTurk workers.

With any further comment, feedback or remark please feel free to email us at socialpsychology.research@yahoo.com.

Appendix B: Additional studies of Chapter 2

Pilot studies for initial test of hypothesis

The pilot studies (1 and 2) were conducted in the US among Whites with African Americans as the target of prejudice, and study materials reflected beliefs about African Americans' intellectual and cognitive skills (e.g., Ashburn-Nardo & Johnson, 2008; Aronson, Fried, & Good, 2002). Participants observed the Logic-IQ game (which we designed for the current research, and pre-tested) and were randomly assigned to either witness a player being prejudiced against African Americans (with an opportunity to respond to the prejudiced player; prejudice condition) or did not witness prejudice (no prejudice condition). In pilot study 2, we had an additional, exposure condition, where participants witnessed the same prejudice but had no opportunity to confront (like in Study 3 in the manuscript).

We measured outgroup attitudes by assessing explicitly perceived abilities of African Americans (rational, competent, intelligent [only in pilot study 2]). We also measured monetary support for a Black organization promoting education (as done similarly in the field; e.g., Freeman, Aquino, & McFerran, 2009; Reimer et al., 2017). Social desirability bias and fear of appearing prejudiced (Crandall, Eshelman, & O'Brien, 2002; Norton, Sommers, Apfelbaum, Pura, & Ariely, 2006; Shelton, West, & Trail, 2010) likely affected participants' responses, and we found no significant differences on perceived abilities (*p*'s > .25), therefore already in pilot Study 2, we mainly focused on the support measure. This measure meant to capture attitudes and intentions towards the outgroup with tapping into prejudicial notions about intellectual abilities (Ashburn-Nardo & Johnson, 2008; Aronson, Fried, & Good, 2002; Devine & Elliot, 1995; Steele & Aronson, 1995), yet it is more subtle and does not require participants explicitly stating how they perceive Black people. In addition, participants have a stake at responding (since we ask them about donating their money) and thus potentially elicit more candid responses.

To ensure that individual socio-political differences do not account for our potential findings, we assessed and controlled for participants' baseline socio-political–intergroup orientations (IMS, SDO in pilot study 1, and political ideology [conservative-liberal, democrat vs. republican] and SJ in both).

Pre-test for Logic-IQ game paradigm

The aim of the pre-test was to establish that the game indeed manipulates perceived racism and that it is credible. In the pre-test, 49 White mTurk workers (71.4% female, M_{age}=29.04 years, SD_{age}=8.99) participated in an online study for monetary compensation (50 cent) and were randomly assigned to an exposure (n=25) and no prejudice condition (n=24). For the pre-test participants had no opportunity to confront the prejudiced player (to avoid non-confronters' justification outcomes). Participants observed the Logic-IQ game and following the game they were asked multiple close-ended questions about the scenario to assess whether they believed they had witnessed racism, what they had witnessed, and whether the game in general was credible. First, we asked participants: “Did you witness something that upset you during the game?”, where answer options were ‘Yes’ or ‘No’. In the exposure condition, 79% responded with ‘Yes’ and in the no prejudice condition only 4%. These differences were significant, $\chi(1)=27.77$, $p<.001$. Our next question was: “Did you witness anything insulting?” (Yes or No). In the exposure condition 75% responded with ‘Yes’ and in the no prejudice condition only 12%. These differences were significant, $\chi(1) =19.05$, $p<.001$. Then we informed participants and asked: “We asked you this, because a participant has indicated in a comment that one of the picker players has been showing signs of negative racial discrimination (dropping out Black players). Did you witness such behavior?” (Yes or No). In the exposure condition 100% responded with ‘Yes’ and in the no prejudice condition nobody, $\chi(1)=48.00$, $p<.001$. Finally, we asked “Did you witness the followings? Discrimination, Racist remark, Both, or None of the above”. In the exposure condition 72% reported to witnessed both, 24% reported to witness at least one of them and the remaining 4% (n=1) said to witnessed neither. In the no prejudice condition, 12% (n=3) claimed to witness both, 4% (n=1) said to witness at least one, and the remaining 84% reported not to witness neither. The overall model was significant, $\chi(3)=32.89$, $p<.001$.

Next, we asked questions from participants in the exposure condition on their assessment of the racist incident: whether they perceived it as racism (*racism scale*), whether it was upsetting or shocking (*upsetting scale*), whether they thought the person was not actually serious (*joking scale*). We asked to what extent they agree with the following items regarding the statement and behavior that has occurred (on a continuous scale from 0 ‘not at all’ to 10 ‘very much’): (1) What I've seen was racial discrimination, (2) What I've seen was racism, (3) His statement was very serious, (4) The statement wasn't that upsetting (reversed), (5) I was shocked by the statement, (6) I thought that he (the picker player) was only joking, (7) I thought that he is not serious. Mean scores on item 1 and 2 were calculated for each participant composing a reliable racism scale ($r=.71$). Mean scores on items 3, 4 and 5 were averaged into the upsetting scale ($\alpha=.56$), and items 6 and 7 composed the joking scale ($r=.72$). Descriptive results indicate that participants scored much higher than mid-point of the scale (‘5’) on the racism measure ($M=8.10$, $SD=1.67$), somewhat higher than mid-point on the upsetting scale ($M=6.58$, $SD=2.06$), and much lower than mid-point on the joking scale ($M=1.54$, $SD=2.47$). Overall, these results allow us to conclude that the Logic-IQ game manipulates racism.

Regarding the credibility of the stimulus, we asked all participants on a scale from 0 (not at all) to 10 (very much) whether the “game looked…”: *like a video*, *fake*, *unrealistic*. Even considering that this is a leading question and it appeared after all the above questions, mean scores on these variables were not particularly high (video:

M=3.84, *SD*=3.61; fake: *M*=6.41, *SD*=3.36, unrealistic: *M*=5.78, *SD*=3.71). There were no significant differences between the exposure and no prejudice condition on these questions, *p*'s>.18.

Overall, we found that a substantial majority of participants who were exposed to racism indicated to have witnessed some form of racism. They also generally thought it was upsetting and they did not think it was a joke. Participants' further responses suggested that the game scenario was credible (did not seem fake). Nevertheless, in the following experiments, we tested and screened participants who figured out the purpose of the study.

See scenes from the game in the manuscript.

Pilot study 1

Methods.

Participants. We recruited mTurk U.S. participants (in 2015 summer) who were asked to participate in an online study for monetary compensation ($2.20). We ran a trial with one participant and then opened the survey for 120 participants. Participants first filled out demographic questions. Respondents who identified as Black/African American/Afro-Caribbean or Latino/Hispanic at the beginning of the survey were directed to the no prejudice condition and their data were dropped from analyses (n=22). The remaining 99 participants were randomly assigned either to the prejudice or no prejudice group. We excluded from analyses those who failed the attention check questions (n=10) and those who figured out the purpose of the study (n=7), leaving 82 participants in total (n=38 in prejudice, n=44 in no prejudice condition; 59.8% female, 39% male, 1.2% other, M_{age}=32.24 years, SD_{age}=10.28).

To conceal the purpose of the study, participants were told we test "how the presence of others facilitates or impairs performance", and that their role of observers serve to increase the players' feeling of being "exposed" during performance. Political ideology (conservative-liberal, democrat vs. republican) was assessed in the demographics. Following the game and filler questions about the game, a seemingly unrelated "sociology" survey about social issues appeared also with filler questions. In this section we included the scales of SJ, SDO, IMS-EMS. Importantly, they responded to our outgroup attitude measure, where we assessed monetary support offered to a Black organization. To increase validity and significance of our dependent measure and participants' perceived (real) stake to responding, we incorporated a real ongoing fundraising campaign, and we did not clarify that this is a hypothetical question. Finally, we asked about participants' policy agreements.

At the end of the study, in order to assess trivialization, we told participants that a survey respondent reported about a possibly racist player. In the prejudice condition, we were vague in this description (in order to decrease suspicion) and asked them if they encountered such behavior; while in the no prejudice condition we described the racist situation exactly as it occurred in other conditions (see manuscript for script). Then we included the trivialization scale (referring to the prejudiced event in all conditions). Finally, participants were debriefed (see manuscript for script).

Stimuli. Participants were first provided with description of the study and game, where we included elements aimed to increase the credibility of our cover story. The game was described as composing of four players (each appearing with a pictogram or photo). The goal was to win as much money as possible by answering logic questions correctly as a group. In each group of players there is a designated established player (called the "Picker"), who eliminates one player at the end of each round (who then loses all his earned money). After elimination, all points (money) are divided equally between the three remaining players (thus, it is beneficial to keep better players and drop the weakest), and then the observer would allegedly join the game. Following instructions and "training", participants were directed to a seemingly different online surface, where in reality, they were exposed to a pre-recorded video (see Figure 1 in manuscript for scenes).

All participants observed the game of three White players and one Black player (see Figure 1a). During the game, participants received a decoy message from the Picker (aimed to decrease suspicion and to draw attention to the screen). After the last question, a performance sheet appeared with players and their earned points. To manipulate prejudice and discrimination, in the prejudice conditions the Picker player eliminated the Black player (see Figure 1b), who was not the weakest player, and then privately messaged the participant saying "He will lose us points. They are not good at these things… You know" (see Figure 1c). In the no prejudice condition, the Picker player dropped the actual weakest (White) player and privately messaged the observer saying "He will lose us points. He is not performing well at all… You saw". Following this message, for all participants, a text box appeared where they could message the Picker player (i.e., confront for his prejudiced behavior). After some delay, a system error message appeared and the game ended.

Note, we told participants at the beginning of the game that when they play, they will play with same Picker player they see in the observing phase. This was done in order to create some ramifications for confronting, like getting retaliation/eliminated, which are otherwise present in face-to-face interaction (e.g., Shelton & Stewart, 2004).

Measures. We measured SJ, SDO, IMS-EMS (see Appendix A). Other measures are included here.

Perceived abilities. We assessed explicit perceptions about African Americans with two items. Participants had to indicate on a slider from 0 (not at all) to 10 (very much) the extent they see "the average member of the following groups as "competent", and "rational". Among the groups, we included African Americans. There was low correlation between these items (r = .41) thus we analyzed them independently.

Support to Black organization. To assess attitudes towards Black people, we measured monetary support offered to a Black tech education program. Participants were asked to read a real and active campaign on

Indiegogo (crowd-fundraising website), titled “Grow Ferguson’s Youth Tech Impact Program”. We inserted a screenshot of the campaign, its shortened description and its page link. Participants read about a predominantly Black organization, which is raising money for their tech impact program (“a six-week intensive program aimed at teaching web development to St. Louis residents between the ages of 16 and 30, with the goal of strengthening local, black and brown-owned businesses and nonprofits in the area. At the end of the program, participants receive a $500 stipend and a laptop (valued at $700) to continue their work in the community”). Under the text, we asked participants “How much of your own monetary reward for this HIT [survey] would you be willing to donate to this project? Please indicate your answer in the form of a percentage.”, they had to respond on a slider from 0 to 100, the slider side-tag read “I’m willing to donate __ % of my reward.” On purpose, we did not clarify that this is a hypothetical question, so respondents feel their response holds real stakes.

Trivialization. We measured participants’ judgment of the severity of the incident using seven items that we developed for the purpose of the present research. Participants were asked to indicate on a slider ranging from 0 = not at all to 100 = very much: To what extent do you agree with the following regarding the statement and behavior of this picker player we have just described to you? [in the control condition] / that has occurred during the game [in the manipulation condition]? (1) His statement was very serious. (reversed-scored) (2) The statement shocked me. (reversed-scored) (3) I think/thought maybe he (the picker player) was only joking. (4) This is/ What I’ve seen was racial discrimination. (reversed-scored) (5) This is racism. / What I've seen was racism. (reversed-scored) (6) I think he was not serious. (7) The statement wasn't that upsetting. A mean severity perception score was calculated for each participant by collapsing across the seven items (α=.75). Higher score indicated trivializing the racist incident.

Results. We coded the responses to the prejudiced message in the prejudice condition and found that 29% of participants questioned or reproached the player for his behavior and/or comment, that is, confronted (n=11), and the remaining participants in this condition did not (71%, n = 27). We found that there was no significant difference on general prejudice orientations (IMS-EMS, SDO; *p*’s>.25), on socio-political orientation (SJ, conservative-liberal dimension, democrat vs. republican; *p*’s>.25), or on basic demographics (age, education level, relative economic status, gender: female vs. male; *p*’s>.25). This enabled us to perform between-subject analyses on the outcome measures.

Using independent samples t-test, we found the hypothesized effect, participants who did not confront prejudice offered significantly less *donation to a Black tech program* (*M*=5.37, *SD*=11.32) than those in the no prejudice condition (*M*=14.66, *SD*=23.42), *t*(66.12)=2.24, *p*=.029, Cohen’s *d*=0.51 (Levene’s test of homogeneity of variance was significant indicating that the assumption of homogeneity of variance was violated, therefore corrected values were reported.) The difference on monetary support remained significant after controlling for IMS, EMS, SDO, SJ and conservative-liberal, *F*(1,64)=4.07, *p*=.048, η^2=0.06. There were no significant differences between groups on *perceived competence* (*p*>.25) or *perceived rationality* (*p*>.25). We found no significant difference on the *trivialization scale* (*p*>.25).

Discussion. Our results in this Pilot study 1 provide evidence for the validity of our paradigm and support our hypothesis regarding increased negative outgroup attitudes following non-confronting. This study also directed us in how to design pilot study 2. The null-results on the perceived abilities measure might have been due to social desirability concerns, because this scale was more obvious than the support measure and participants also had no stake at responding to it honestly. It also had low internal consistency. We improved this scale by adding an additional item ("intelligent"), but this was considered an additional measure and our main outcome variable was the monetary support measure (because the Indiegogo campaign ended by then, we used a different campaign). In addition, our self-developed measure of trivialization did not reveal significant effects. In pilot study 2 we used a validated scale adapted from the literature to measure trivialization. Finally, to shorten the survey, and decrease suspicion about the nature of the research hypothesis, we did not include the IMS-EMS and SDO scales (due to an omission error, we also took out attention check questions).

Pilot study 2

Methods.

Participants and procedure. We recruited 320 U.S. residents through mTurk (ending up with two extra respondents) in 2015 winter, who participated in our study for monetary compensation ($2.20). We performed power analysis based on the effect size obtained in the pilot study to determine the required sample size to achieve a power of 0.80 for the predicted effects. The calculation indicated that 180 participants are required, but we recruited more anticipating exclusion based on selection strategies. Participants first filled out demographics (see Appendix A for questions across studies). Those who did not identify as White/Caucasian (n=98) were assigned to no prejudice condition and were not analyzed, and the remaining 224 White participants were randomly assigned to one of three conditions: prejudice and opportunity to confront ('prejudice'), prejudice and no opportunity to confront ('exposure') and control with no prejudice. We excluded from analyses those who figured out the purpose of the study (n=11; We asked what was the game about, and excluded participants who wrote "reacting"/"responding"/"intervening"/"being a bystander" to "racism"/"prejudice"/"discrimination", or "standing up for others".), leaving 213 participants in total (n=90 in prejudice, n=63 in exposure, n=60 in no prejudice condition; 55.4% female, 44.1% male, 0.5% other; M_{age}=30.92 years, SD_{age}=9.37). (With using 'Reset Element Count' in Qualtrics we recruited more participants to the prejudice condition to ensure that even after exclusion of those who confronted, number of participants will be equally balanced across conditions.)

Stimuli and procedure. We used the same study protocol and game stimuli as described in pilot study 1, except for the additional exposure condition. In this control condition, participants witnessed exactly the same prejudice and discrimination as in the prejudice condition, however following the message, they did not receive the text box, where they could message the Picker player (i.e., confront for his prejudiced behavior). Like in pilot study 1, confronting in the prejudice condition were determined by reading and coding participants' messages to the racist player. Following the game, participants filled out a seemingly unrelated social survey that included SJ and the outgroup attitude measures (support to Black organization). We measured trivialization with a different scale than in pilot study 1. At the end of the survey, we asked what the study was about.

Measures.

Confronting. We read and coded participants' responses in the message box, which was provided to them following the prejudiced player's private message. In the prejudice condition, those who questioned or reproached the prejudiced player for eliminating the Black player and/or making that remark about him/his group were coded as "confronting" and those who did not question or reproach him were coded as "not confronting".

Support to Black organization. To measure outgroup attitudes, we asked participants to express their willingness to donate to a program that offers STEM education for Black urban youth. Participants were asked to read a short text from Black Lives Matter website that informed participants that donations are used for two purposes: to support social justice grassroots organizations and to support S.T.E.A.M. learning for urban youth. For the latter the description was: "The BLCKBOX Campaign is a monthly subscription box of enriching activities for urban youth to experience, the box is based on S.T.E.A.M (science, technology, engineering, arts and math) tools. Because of the huge disparity with black children and S.T.E.A.M and with S.T.E.A.M providing the financial opportunities to security and economic parity to the Black communities and families. Our goal is to provide 100,000 boxes quarterly to black children ages 7 – 11." Participants were requested to indicate "How much money between 0 to $50 would you be willing to donate to this organization" on a slider from 0 to 50.

Trivialization. Same as in Studies 1–3.

Results. First, we coded responses and determined those who did not confront the prejudiced player (n=60, around 67%; meanwhile n=30 confronted). The study groups (non-confronters, exposure and no prejudice) did not differ on demographics (gender, education or relative economic situation; *p*'s>.25, age *p*<.05) or socio-political attitudes (SJ, conservative-liberal, and democrat vs. republican; *p*'s>.25). This enabled us to perform between-subjects analyses on the outcome measures.

To test our main prediction, we performed Univariate ANOVA with contrast test and controlled for political ideology (conservative-liberal) and SJ. Omnibus ANOVA was significant, *F*(2,178)=3.62, *p*=.03, η^2=.04. As predicted, we found that White participants who did not confront prejudice albeit having opportunity subsequently offered less support to a Black organization (*M*=6.65, *SD*=8.73) compared to those in the no prejudice condition (*M*=13.10, *SD*=17.09; *p*=.02, *95% CI* [-11.35, -1.18]), or those in exposure condition (*M*=12.71, *SD*=15.47; *p*=.03, *95% CI* [-10.70, -.64]). When not controlling for political ideology and SJ the results remained significant, *p*'s < .02.

We found similar pattern of results for trivialization. Omnibus ANOVA was significant, *F*(2,178)=9.07, *p*<.001, η^2=.09. The non-confronting group perceived the incident as significantly more trivial (*M*=4.06, *SD*=2.00) than those in the no prejudice condition (*M*=3.07, *SD*=1.80); *p*<.01, *95% CI* [.33, 1.61]), or in the exposure condition (*M*=2.67, *SD*=1.63; *p*<.001, *95% CI* [.70, 1.96]). When not controlling for political ideology and SJ these comparisons remain significant, *p*'s<.01.

For **discussion** of these results see manuscript.

Additional measures and analyses. As mentioned in the introduction part, there was no significant effect on perceived abilities (*p*'s>.25).

Pilot test for Trust Game paradigm (Muslim outgroup version)

To pilot the trust game paradigm, prior to the within-subjects studies (Study 1 and 2), we ran an experiment in the US (N=275). In this study, we had three between-subjects conditions: witnessing intergroup prejudice and opportunity to confront (intergroup) vs. witnessing interpersonal prejudice with opportunity to confront (interpersonal) vs. witnessing intergroup prejudice but no opportunity (exposure).

Participants played the Trust game and then responded to the social closeness toward Muslims scale, IMS-EMS, SDO, SJ, trivialization, socio-political orientations, demographics. Unfortunately, results indicated some significant differences on political affiliation (Clinton-Trump voting), IMS and gender between intergroup prejudice non-confronting group vs. the interpersonal non-confronting group and the exposure condition. This precluded us from using the results of this study. We ran this particular study around the US elections in 2016 when anti-Muslim attitudes were a highly divisive issue, which may have affected who was more likely to confront. In the same time, we assessed all these variables *after* the manipulation and not prior due to our concern that they may reveal the study purpose, which however also involved the risk that the experimental manipulation may affect it. Indeed, it is not possible to tell whether the difference on IMS was a difference prior to the experiment, or such attitudes were affected by the experimental manipulation.

Nevertheless, at the end of the survey we tested and analyzed whether we indeed succeeded in manipulating intergroup prejudice in this paradigm and whether the game's credibility did not differ across conditions, thus the issue of demand characteristics could not generally explain potential findings in this paradigm. First, we asked participants whether they witnessed anything insulting during the game (with Yes or No response options). In the interpersonal condition 31.5%, in the intergroup condition 94.5% and in the exposure condition 90.2% responded ‘yes’ (vs. ‘no’), χ(275)=112.78, *p*<.001. Next, in the intergroup and exposure condition we asked: “We asked if you had witnessed anything insulting because a respondent informed us about a player commenting about another player's ethnicity. Did you encounter such behavior? Specifically, did you witness any of the following? (multiple answers possible) – Discrimination, Racist remark/s, None of the above.” In the exposure condition, 96.7% said at least one of them and in the intergroup condition, 98.7%, χ(170)=0.76, *p*>.25. Only 3.3% and 1.3% chose ‘none’, respectively. This question was not asked in the interpersonal condition. Finally, we asked a close-ended question: “Did you think that perhaps the game you observed is not real?” with Yes, Maybe, and No options, and we did not find significant difference between conditions, χ(273)=2.18, *p*>.25.

Appendix C: Additional measures of Chapter 2

List of all measures at post-test in order of appearance:

Study 1: *Outgroup opponent* (see A) – perceived trustworthiness (see manuscript) – trivialization (see manuscript) – denial of responsibility (see manuscript) – *confronting intentions* (see D)

Study 2: Outgroup opponent (see A) – social closeness (see manuscript) – *policy agreement* (see “B”) – trivialization (see manuscript) – denial of responsibility (see manuscript) – confronting intentions (see D)

Study 3: Feeling thermometer (see manuscript) – MRS (see manuscript) – trivialization (see manuscript) – denial of responsibility (see manuscript) – confronting intentions (see D)

A. *Trusting an outgroup opponent in the game (Study 1 and 2)*

As an additional secondary outgroup attitude measure, we wanted to see whether participants will be less trusting of an outgroup member as an effect of witnessing, and not confronting racism. This was not an outcome measure of main focus because we wanted to see attitudes towards the whole group, and not solely an outgroup individual, and because this could not be pre-tested (assessed prior to the game).

In Study 1 and 2, during the end of the Trust Game paradigm participants could play the game themselves and their second opponent was a player with a nickname of ‘Klezmer50’ (Study 1) or ‘Salim’ (Study 2), and as with other opponents, participants could choose to give him all/half/none. The system broke down right after they had made a choice. Chi-square test of independence showed no significant effects on this measure (p’s>.25; tested between intergroup and interpersonal non-confronting). In the end of the survey, we asked participants whether they thought this player was Jewish/Muslim (‘yes’ or ‘no’), and even with such a leading question, only 8% (n=15, study 1) and 27% (n=33, study 2) of the participants indicated that they thought he may be Jewish/Muslim. A sample including only those who understood he may have been Jewish/Muslim were too small for analyses. We did not choose more obvious names for the outgroup players because we were concerned it may expose the purpose of the study and may influence responses on the subsequent main outgroup attitude measures. Due to this assessment challenge, we did not include such a measure in study 3.

B. *Outgroup-related policy agreement (Study 2)*

Data collection occurred during the Muslim ban in the U.S. therefore for exploratory purposes, we included a scale of agreement to political policies directed at Muslims. While we predicted that intergroup attitudes corresponding to the prejudicial incident would change, we were uncertain whether this effect would transfer to policy support, given that they are less directly relevant to the witnessed event.

To this purpose we asked participants their agreement (on a 6-point scale from 1 = completely disagree to 6 = completely agree) to the following items (*adapted and altered from Kteily & Bruneau, 2107; Lee et al., 2013; Oswald, 2005)*: 1. We should not accept Muslim immigrants into the U.S. (r). 2. Muslim residents or visitors should leave the country. (r). 3. I would support any policy that would stop the building of new mosques. (r) 4. Head scarves should be banned in all public spaces. (r) 5. There should be careful security checks of Muslims when entering crowded public spaces, like train stations and airports. (r). 6. If a Muslim person is under suspicion, I think the government should be allowed to monitor his online activity. (r) 7. Hate speech against Muslims should be punished by law. 8. Muslims should be submitted to equal and fair treatment in all aspects of life. 9. Muslims are underrepresented in the parliament, and they should be represented more. 10. The media portrayal of Muslims as violent and dangerous people is wrong and it should be regulated. Items were averaged into a policy scale (10 items, α=.91), with *higher scores denoting higher anti-Muslim stance*.

There was no significant *baseline* difference between intergroup non-confronters and interpersonal non-confronters (*p*=.13). The interaction between time and experimental groups on policy agreement was not significant, *F*(1,118)=3.40, *p*=.068. Simple effects analyses indicated no overtime change among intergroup non-confronters (*p*=.17; pre-test: *M*=3.19, *SD*=1.24; post-test: *M*=3.09, *SD*=1.18) or interpersonal non-confronters (*p*=.22; pre-test: *M*=2.83, *SD*=1.28; post-test: *M*=2.90, *SD*=1.22). Further simple effects analysis showed no difference between groups at post-test (*p*>.25). Similarly, while controlling for baseline policy agreement, using between-subjects Univariate ANCOVA, we found no significant difference between conditions (*p*=.16).

C. Confronting intentions

We asked some questions about participants' intention to confront the witnessed/told incident both in the prejudice and control conditions for exploratory purposes. It also provided us with some indication of confronting norm in each intergroup context, as reflecting in the hypothetical confronting scores in the control conditions (as mentioned in the general discussion).

[in prejudice with opportunity conditions]:

– "Did you confront the Picker [racist] player in the message at the end of the observing session? If yes, to what extent?" (Pilot studies) / "Did you confront the mentioned player during the game and if yes, to what extent?" (S1 and S2) on a slider from 0 = I did not confront him to 100 = I totally confronted him / "Did you confront the player for his/her behavior toward the minority individual?" 1=Yes, I confronted the player for his/her behavior., 0 = No, I did not confront the player for his/her behavior. (S3)

- “To what extent did you consider confronting that player?” on a slider from 0 = I did not consider confronting at all to 100 = I considered confronting very much (not in Study 3)
- “Do you think you should have confronted the Picker player/that player?” (not in Study 3)
- “Do you feel regret about not confronting this behavior?” (not in Study 3)

[in control conditions] Hypothetical confronting estimation*:

- (in Pilot studies, S1-S2) “If you would have witnessed this incident, do you think you would have confronted the mentioned player by messaging him?” on a slider from 0 = I would’ve not confronted him to 100 = I would’ve totally confronted him.
- (S3, exposure/interpersonal condition) “If you had an opportunity to message him, do you think you would have confronted the player for his behavior toward the minority individual?” / “If you would have witnessed this incident and had an opportunity to message him, do you think you would have confronted the player for his behavior toward the minority individual?” 1=Yes, I would have confronted the player for his behavior. 0 = No, I would have not confront the player for his behavior.

*The *means* (and *SD’s*) across studies on this question were: *51.01* (30.42) in Study 1 and *50.86* (33.04) in Study 2. In Study 3, 68/65% (exposure/interpersonal) said they would have confronted the prejudicial event.

Appendix D: Additional analyses of Chapter 2

Power analyses across studies

In study 1, sample size was determined based on availability in the student credit pool. We conducted a sensitivity power analysis in G*Power 3.1 as follows: F tests > ANOVA: repeated measures, within-between interaction > Sensitivity power analysis > Input parameters: alpha=0.05, power=.80, total sample size=138, number of groups=2, number of measurements=2, corr among repeated measures=0.5, nonsphericity correction=1. We found that a mixed ANOVA with 138 participants for two time points and two between factors are sensitive to effects of η^2_p = .01 (exact: Cohen's *f*=.12) with 80% power (alpha = .05).

In study 2, we determined our sample size a-priori considering the moderation analysis with IMS, which however was no longer in the focus of the manuscript. Therefore, we instead calculated a sensitivity power analysis for the mixed model (2 measurement, between-IV with 2 levels) in G*Power 3.1 as follows: F tests > ANOVA: repeated measures, within-between interaction > Sensitivity power analysis > Input parameters: alpha=0.05, power=.80, total sample size=120, number of groups=2, number of measurements=2, corr among repeated measures=0.5, nonsphericity correction=1. We found that a mixed ANOVA with 123 participants for two time points and two between factors would be sensitive to effects of η^2_p = .02 (exact: Cohen's *f*=.13) with 80% power (alpha = .05).

For study 3, using G*Power we calculated that for the moderation analysis, between-subjects IV with 3 levels and a moderator (2 tested predictors, 5 total predictors), aiming 0.20 effect size (*this is how it was mistakenly pre-registered, in fact, we calculated with Cohen's f^2 = .02, making $f \approx .14$*) and 0.80 power, the required sample size is 485. However, as a data-driven post-hoc decision, for this moderation analysis, we collapsed the two control conditions, so the between-IV had 2 levels. To this end, we conducted a sensitivity power analysis in G*Power as follows: F tests > Linear multiple regression: Fixed model, R2 increase > Sensitivity power analysis > Input parameters: alpha=0.05, power=.80, total sample size=410, number of tested predictors=1, total number of predictors=3. Which analysis indicated that the moderation analysis with 410 participants would be sensitive to effects of η^2_p *= .02 (*exact: Cohen's *f^2 = .02;* Cohen's *f* =.14) with 80% power (alpha = .05).

Intergroup prejudice confronters

There was no significant outgroup attitude change among intergroup confronters from pre-test to post-test across studies – **Study 1** (n=8): M_{pre}=5.04, SD_{pre}=0.81, M_{post}=4.25, SD_{post}=0.83, *p*=.17; **Study 2** (n=42): M_{pre}=4.38, SD_{pre}=1.33, M_{post}=4.47, SD_{post}=1.34, *p*=.51; **Study 3** (n=37): M_{pre}=80.76, SD_{pre}=18.78, M_{post}=81.51, SD_{post}=19.01, *p*=.68.

Regarding differences between intergroup confronters and other experimental groups on socio-political orientations, importance of confronting or outgroup prejudice – In **Study 1**: *p*'s > .33; In **Study 2**: *p*'s> .18, except on SDO (*p*=.022 between intergroup confronters vs. intergroup non-confronters); In **Study 3**: *p*'s > .12, except SDO (*p*=.041 between intergroup confronters vs. intergroup non-confronters).

Moderation analysis with perceived importance of confronting (Study 3)

Table S1. Main, interactive and simple effects (and means) for the effect of experimental groups (intergroup non-confronting vs. interpersonal non-confronting vs. exposure) and perceived importance of confronting scale (moderator) on feeling-thermometer (0-100) in Study 3.

		Means					
Predictor	*intergroup non-conf*	*interpers. non-conf*	*exposure*	*B (SE)*	*t*	*p-value*	*95% CI*
Importance	-	-	-	.40 (0.53)	.76	.451	-0.64,1.43
D1 (Intergroup vs. Interpersonal)	-	-	-	-.89 (5.30)	-.17	.867	-11.31,9.53
D2 (Integroup vs. Exposure)	-	-	-	4.62 (5.04)	.92	.360	-5.29,14.53
D1 x Importance	-	-	-	0.65 (0.74)	.87	.383	-0.81,2.10
D2 x Importance	-	-	-	-0.11 (0.70)	-.16	.873	-1.49,1.26
FT prescores	-	-	-	.75 (0.03)	22.62	.000	0.69,0.82
Low-importance	72.07	74.07	76.19	*2.01 (2.43)* 4.12 (2.35) ^{a}	*.83* 1.75	*.410* .081	*-2.78,6.79* -0.50,8.74
Mean-importance	73.01	76.54	76.86	***3.53 (1.74)*** **3.86 (1.72)**	**2.02** **2.24**	***.044*** **.025**	***0.10,6.97*** **0.48,7.23**
High-importance	73.94	79.01	77.53	***5.06 (2.51)*** 3.59 (2.42)	**2.02** 1.49	***.044*** .138	***0.13,9.99*** -1.16,8.34

Note. Perceived importance of confronting was tested on a 10-point scale and conditioning values (simple slopes analyses) were based on 1 SD above, mean, 1 SD below as follows: 4.48 (for low), 6.85 (for mean) and 9.22 (for high). ^{a} D1 on top in italics, D2 on bottom. Significant simple effects in bold.

Appendix E: Materials of Chapter 3

Demographic questions

Lastly, please answer some demographic questions about yourself:

Age
____

_____

Gender

- ○ Male (1)
- ○ Female (2)
- ○ Other (3) ____

Your highest education level:

- ○ Less than high school (1)
- ○ High school diploma (2)
- ○ Bachelor's degree (3)
- ○ Master's degree (4)
- ○ PhD (5)
- ○ Other [not included in analyses]

What is the race or ethnicity which you identify the most with?

- ○ White or Caucasian (1)
- ○ Black/African American/Afro-Caribbean (2)
- ○ Latino/Hispanic (3)
- ○ Native American (4)
- ○ Asian (5)
- ○ Arab (6)
- ○ Biracial / Mixed: (7) ____
- ○ Other (8) ____

What is your religion?

- ○ No religious affiliation (1)
- ○ Christian (2)
- ○ Muslim (3)
- ○ Jewish (4)
- ○ Hindu (5)
- ○ Buddhist (6)
- ○ Other (7)

What is your socio-political orientation?
Conservative
Liberal

0 1 2 3 4 5 6 7 8
9 10

continuous slider

What is your political affiliation?

- o Democrat (1)
- o Republican (2)
- o Neither (3)
- o Don't want to answer (4)

Compared to other people in your society, what is your economic situation?

- o Wealthy (1)
- o Better than most (2)
- o Good (3)
- o So-So (4)
- o Poor (5)
- o Destitute (6)

coding reversed in study analyses

Moral-prejudice identity self-importance

(adapted from Aquino & Reed, 2002)

Think about a person who is not prejudiced, who is egalitarian and believes that all people are created equal, and who does not discriminate against people based on their gender, sexual orientation, ethnicity or religion. It could be you or it could be someone else. For a moment, visualize in your mind the kind of person who has these characteristics. Imagine how that person would think, feel, and act. When you have a clear image of what this person would be like, answer the following questions.
(9-point scale from 1 = not at all true of me to 9 = completely true of me)

[Internalization scale]

1. It would make me feel good to be a person who has these views and beliefs
2. Being someone who has these views and beliefs is an important part of who I am.
3. I would be ashamed to be a person who has these views and beliefs. (reversed-scored)
4. Having these views and beliefs is not really important to me. (reversed-scored)
5. I strongly desire to have these views and beliefs.

[Symbolization scale]

6. The types of things I do in my spare time (e.g., hobbies) clearly identify me as having these views and believes.
7. The kinds of books and magazines that I read identify me as having these views and beliefs.

8. The fact that I have these views and beliefs is communicated to others by my membership in certain organizations.
9. I am actively involved in activities that communicate to others that I have these views and beliefs.

From the original scale we excluded the following item: "I often wear clothes that identify me as having these characteristics" (symbolization item).

Vignette scenarios and confronting intentions (in Study 4)

In this part of the survey, you'll be presented with two ambiguous situations that pose moral dilemmas and you'll be asked questions about it. Please try to place yourself in those described situations as much as possible. Please respond to the questions honestly, according to your own belief, and not according to what you think is expected of you.

[Scenario A]

Imagine you are traveling on the bus, sitting in the back. It's a big bus, only a few people traveling on it. An older teenage boy boards the bus, sits at the back, not far from you, and you hear that he is speaking in Spanish on the phone. Near him, a middle-aged man is sitting, who keeps staring at the boy. The boy hangs up the phone, and the man starts to speak to him. He tells the boy this is the US, and people speak English here. He continues and says that immigrants like him [the boy] should leave this country. Because you believe that this specific boy is treated unfairly, you are debating whether to intervene or not. On the one hand, if you get involved, the man may verbally or even physically attack you. You also don't want to miss your stop which you are approaching soon. If you miss your stop, you'll be late for an important appointment. [**All participants read the prior part. Those in the control condition stopped here**] [**Moral loss framing condition continued with the following text:**] On the other hand, if you don't get involved, you will probably feel like a bad person. You believe that this action would reveal a bad side of you. That is, you feel that in this situation staying silent means you are behaving immorally. You keep thinking that if you want to avoid moral failure, you should probably intervene. [**Moral gain framing condition continued with the following text:**] On the other hand, if you get involved, you will probably feel like a good person. You believe that this action would reveal a good side of you. That is, you feel that behaving morally in this situation means speaking up. You keep thinking that if you want to fulfill your moral ideals, you should intervene.

[Scenario A – Questions]

Based on solely what is described in the text, considering the potential risks involved, how likely it is that in this situation you would perform the following behaviors? (1 = not likely at all to 9 = very much likely)

1. I would stay in my seat and I would not get involved. [reversed]
2. I would quietly leave them and move to the front of the bus. [reversed]
3. I would confront the man and tell him he is racist.
4. I would ask the man to stop assaulting the boy.
5. I would sit next to the boy and start talking to the boy in a friendly manner.
6. I would ask the bus driver to stop the man's behavior.
7. Other suggestion (not mandatory, if you don't write, just mark 1): _________

Overall, to what extent you would confront in this situation in order to help the boy?
1 = I would not confront at all to 9 = I would totally confront on a 9-point scale

[Scenario B]

Imagine you are at work, sitting at your desk, working on a difficult assignment. Then you slowly become attentive to a conversation that two of your co-workers are having in the adjacent room. You hear them talking about a third co-worker, who is Muslim. You hear them laughing and making fun of her headscarf, making nasty references about her because of her religion. You don't have many feelings about the mentioned Muslim co-worker, because you hardly know her, but you also don't think it's nice to talk about another individual like that. So, you are debating to confront your co-workers or not. On the one hand, you don't want them to think that you often eavesdrop on their conversations. Additionally, you are working closely with these colleagues, and if they get offended they could even jeopardize your position at work. You also must finish the assignment you are working on as soon as possible. [**All participants read the prior part. Those in the control condition stopped here**] [**Moral loss framing condition continued with the following text:**] On the other hand, you are now recalling other unfair situations you've witnessed in the past and how badly you felt about yourself after not confronting. You feel it is your moral obligation to intervene. If you don't intervene, you fail your moral duty, and you may later feel like a worse person morally. You feel you can lose a lot if you don't confront. [**Moral gain framing condition continued with the following text:**] On the other hand, you are now recalling other unfair situations you've witnessed in the past and how better you felt about yourself after confronting. You feel it is your moral aspiration to intervene. If you intervene, you succeed to live up to your moral principles, and you may later feel like a better person morally. You feel you can gain a lot if you confront.

[Scenario B – Questions]

Based on solely what is described in the text, considering the potential risks involved, how likely it is that in this situation you would perform the following behaviors? (1 = not likely at all to 9 = very much likely)

1. I would stay in my office and I would not confront them. [reversed]
2. I would sit somewhere else so I don't hear them but I would not confront them. [reversed]
3. I would ask my co-workers to stop insulting her.
4. I would tell my supervisor about my co-workers' conversation.
5. Without being specific, I would just ask them to keep quiet while making sure they know I disprove of their conversation.
6. I would confront my co-workers and tell them they are racists.
7. other suggestion (not mandatory, if you don't write, just mark 1): __________

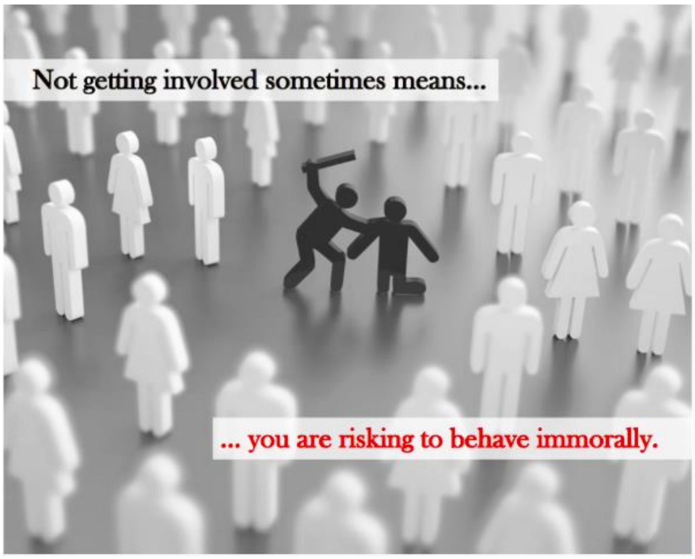
Overall, to what extent you would confront in this situation in order to stand up for her?
1 = I would not confront at all to 9 = I would totally confront on a 9-point scale

Moral mindset intervention (in Study 5)

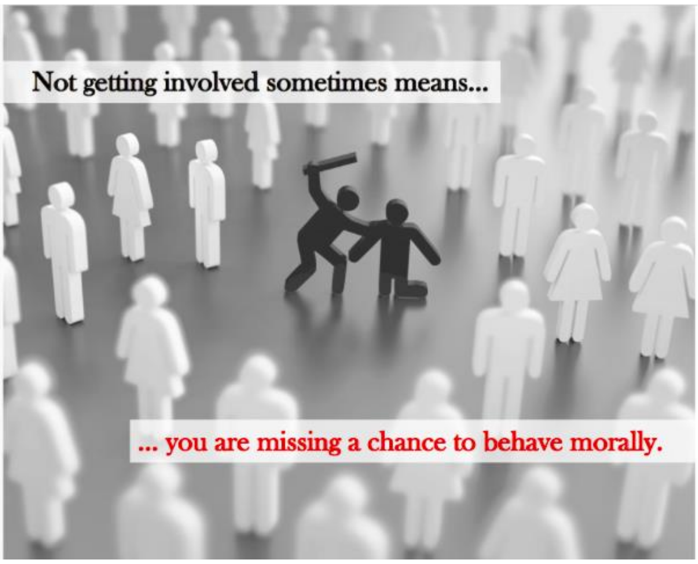
Letters are bolded to emphasize the differences between conditions.

[Task 1]

[Loss condition]


Not getting involved sometimes means...
... you are risking to behave immorally.

[Gain condition]


Not getting involved sometimes means...
... you are missing a chance to behave morally.

Please describe what you think this poster means. (please write min. 350 characters)

[Task 2]

Please see a video that was featured in the local news depicting a British woman insulting other passengers she assumed to be Polish or immigrants on a public bus: [Here appeared a video]

[Loss condition]

Later when this incident was reported in the news, a passenger on this bus said that he wished he would have stopped the woman from insulting those other passengers.

Imagine you are this bystander who did not intervene. Please give a short account of your thoughts and feelings.

Start with **“I feel like not intervening revealed a bad side of me…”**

(please write min. 350 characters)

[Gain condition]

Later when this incident was reported in the news, a passenger claimed that he wished he would have stopped the woman from insulting those other passengers.

Imagine you are this bystander who did not intervene. Please give a short account of your thoughts and feelings.

Start with **“I feel like intervening would have revealed a good side of me…”**

(please write min. 350 characters)

[Task 3]

[Loss condition]

Stanislaw Chmielewski (1909–1992) is a Polish Christian man who risked his life to confront injustice and saved a dozen Jews during the Holocaust. He once noted that **not doing what he did would have cost him his moral virtue and he would have felt like a bad person**.

Without knowing more about him, how would you describe Stanislaw's potential thoughts and feelings about his own behavior? (please write min. 350 characters)

[Gain condition]

Stanislaw Chmielewski (1909–1992) is a Polish Christian man who risked his life to confront injustice and saved a dozen Jews during the Holocaust. He once noted that **through this action he gained moral virtue and he feels he became a better person for doing it**.

Without knowing more about him, how would you describe Stanislaw's potential thoughts and feelings about his own behavior? (please write min. 350 characters)

Appendix F: Additional measures of Chapter 3

IMS-EMS and SDO

For exploratory and future research purposed we included a Social Dominance Orientation scale (SDO; Pratto et al., 2013) and Internal/External Motivation to Respond Without Prejudice Scale (IMS/EMS; Plant & Devine, 1998). See below scales, which were placed after the morality scales in both studies.

IMS–EMS

To what extent do you agree or disagree with the following statements:
(9-point scale from 1 = not at all true of me to 9 = completely true of me)

1. Because of today's PC (politically correct) standards I try to appear nonprejudiced towards different groups.
2. I try to hide any negative thoughts about other groups in order to avoid negative reactions from others.
3. If I acted prejudiced toward other groups, I would be concerned that others would be angry with me.
4. I attempt to appear nonprejudiced toward other groups in order to avoid disapproval from others.
5. I try to act nonprejudiced toward other groups because of pressure from others.
6. I attempt to act in nonprejudiced ways toward other groups because it is personally important to me.
7. According to my personal values, using stereotypes about other groups is OK. (reversed-scored)
8. I am personally motivated by my beliefs to be nonprejudiced toward other groups.
9. Because of my personal values, I believe that using stereotypes about other groups is wrong.
10. Being nonprejudiced toward other groups is important to my self-concept.

Items 1 through 5 are EMS. Items 6 through 10 are IMS.

SDO

There are many kinds of groups in the world: men and women, ethnic and religious group, nationalities, political fractions. How much do you support or oppose the ideas about groups in general?
(9-point scale from 1 = extremely oppose to 9 = extremely favor)

1. In setting priorities, we must consider all groups.
2. We should not push for group equality.
3. Group equality should be our ideal. (reversed-scored)
4. Superior groups should dominate inferior groups.

Appendix G: Additional analyses of Chapter 3

General moral ought vs. ideal orientation

The moral ought-ideal scale (Aoki, 2015) was placed in the same block as the moral commitment to non-prejudice scales in both studies. Please see the scale below:

To what extent these statements are true or not about the goals you have regarding moral behavior
(9-point scale from 1 = not at all true of me to 9 = completely true of me)

1. *I often feel like my moral goals are things I have to do.*
2. *I am often anxious about falling short of my moral responsibilities.*
3. *A major goal in my life is to avoid moral failures.*
4. *I often think about how I might fail to be a morally good person.*
5. *I worry about becoming a worse person morally.*
6. *Morally, I see myself as someone who is striving to reach my ideals – to fulfill my moral hopes and aspirations.*
7. *I often feel like my moral goals go beyond what is simply required.*
8. *Achieving my moral goals is a personal choice, not simply a duty.*
9. *I generally focus on how I will achieve my moral aspirations.*
10. *Going beyond what is minimally required to be a moral person is very important to me.*
11. *A major goal in my life right now is to achieve my moral ambitions.*
12. *I focus on how I can become a better person morally.*

Items 1 through 5 compose the Moral Ought scale (α = .84 in Study 4 and α = .85 in Study 5) and items 6 through 12 compose the Moral Ideal scale (α = .90 in Study 4 and α = .92 in Study 5). Below are the correlations between the moral ought-ideal scales and study variables:

	M* *(SD)	**Conf. intentions**	**Moral conviction**	**MID-internal**	**MID-symbol**	**Cons–Liberal**	**SES**	**Edu**	**Age**	**Moral ought**
Study 4										
Moral ought	*4.68 (1.97)*	$.15^{**}$	$.21^{**}$	$-.11^{*}$	$.28^{**}$	$-.11^{*}$	$.11^{*}$	.07	$-.18^{**}$	-
Moral ideal	*6.03 (1.84)*	$.29^{**}$	$.53^{**}$	$.23^{**}$	$.47^{**}$	-.08	$.14^{**}$	.05	.06	$.54^{**}$
Study 5										
Moral ought	*4.81 (1.96)*	-.09	.10	$–.19^{**}$	$.21^{**}$	-.001	.11	$.17^{**}$	$–.19^{**}$	-
Moral ideal	*6.14 (1.79)*	-.04	$.42^{**}$	.11	$.39^{**}$	$–.16^{**}$	$.21^{**}$	.08	–.01	$.50^{**}$

Note. * p < .05, ** p < .01 (N = 429 in Study 4 and N = 260 in Study 5)

This scale did not moderate the relationship between moral mindsets and confronting intentions neither in Study 4 or in Study 5, and there were no significant simple effects either.

See the effect of moral mindset condition (control, loss, gain) on confronting intentions as a factor of the scales:

		Confronting			
Moderator	*Predictor*	*B (SE)*	*t / Z*	*p-value*	*95% CI*
Moral ought					
	Moral ought	.14 (.07) **-.09 (.11)**	1.98 **-.78**	.05 **.43**	.00; .27 **-.31; .13**
	D1 (Control vs. Loss)	.48 (.50) **.27 (.86)**	.95 **.32**	.34 **.75**	-.50; 1.45 **-1.40; 1.95**
	D2 (Control vs. Gain)	-.38 (.49) **.36 (.81)**	-.76 **.45**	.45 **.66**	-1.35; .59 **-1.22; 1.94**
	D1 x Moral ought	-.10 (.10) **.02 (.17)**	-1.01 **.11**	.31 **.92**	-.29; .09 **-.31; .35**
	D2 x Moral ought	.06 (.10) **-.05 (.16)**	.57 **-.32**	.57 **.75**	-.13; .25 **-.37; .26**
	Order	-.04 (.17) –	-.22 –	.83 –	-.37; .30 –
Moral ideal					
	Moral ideal	.22 (.07) **-.15 (.11)**	3.11 **-1.31**	.00 **.19**	.08; .36 **-.37; .07**
	D1 (Control vs. Loss)	.05 (.66) **-1.46 (1.23)**	.07 **-1.18**	.94 **.24**	-1.26; 1.35 **-3.86; .95**
	D2 (Control vs. Gain)	-.56 (.62) **-.46 (1.04)**	-.89 **-.44**	.37 **.66**	-1.78; .67 **-2.51; 1.58**
	D1 x Moral ideal	-.01 (.10) **.29 (.19)**	-.07 **1.51**	.95 **.13**	-.21; .20 **-.09; .67**
	D2 x Moral ideal	.09 (.10) **.29 (.19)**	.92 **1.51**	.36 **.13**	-.10; .29 **-.09; .67**
	Order	.00 (.17) –	.00 –	1.00 –	-.33; .33 –

* In the moderation analyses with moral ought orientation, we controlled for moral ideal scale, and vice versa.

** Study 4 results reported on first line; Study 5 results reported on second line in bold.

See simple effects and estimated conditional means for confronting intentions (9-point scale) in Study 4:

	Low on moderator (–1 SD)					High on moderator (+1 SD)				
	Control	Loss	Gain	D1	D2	Control	Loss	Gain	D1	D2
Moral ought	6.78	6.99	6.56	*b* = .21 *SE* = .27 *t* = .76 *p* = .45 [-.33; .74]	*b* = -.23 *SE* = .27 *t* = -.84 *p* = .40 [-.76; .30]	7.32	7.14	7.31	*b* = -.18 *SE* = .27 *t* = -.68 *p* = .50 [-.71; .35]	*b* = -.01 *SE* = .26 *t* = -.04 *p* = .97 [-.53; .52]
Moral ideal	6.62	6.64	6.45	*b* = .02 *SE* = .28 *t* = .07 *p* = .95 [-.52; .56]	*b* = -.17 *SE* = .26 *t* = -.67 *p* = .50 [-.68; .33]	7.44	7.43	7.60	*b* = -.01 *SE* = .25 *t* = -.03 *p* = .98 [-.51; .49]	*b* = .17 *SE* = .27 *t* = .62 *p* = .53 [-.36; .69]

See probabilities (odds in brackets) for each condition and simple effect statistics for confronting action (0 = didn’t confront, 1 = confronted) in Study 5:

	Low on moderator (–1 SD)					High on moderator (+1 SD)				
	Control	Loss	Gain	D1 (Control vs. Loss)	D2 (Control vs. Gain)	Control	Loss	Gain	D1 (Control vs. Loss)	D2 (Control vs. Gain)
Moral ought	.44 (.79)	.52 (1.08)	.50 (1)	*b* = .32 *SE* = .45 *Z* = .72 *p* = .47 [-.55; 1.20] *OR* = .73	*b* = .22 *SE* = .42 *Z* = .50 *p* = .62 [-.63; 1.06] *OR* = .79	.36 (.56)	.46 (.85)	.36 (.56)	*b* = .39 *SE* = .46 *Z* = .86 *p* = .39 [-.50; 1.29] *OR* = .66	*b* = .02 *SE* = .46 *Z* = .04 *p* = .97 [-.88; .91] *OR* = 1
Moral ideal	.47 (.89)	.42 (.72)	.45 (.82)	*b* = -.21 *SE* = .48 *Z* = -.43 *p* = .67 [-1.14; .73] *OR* = 1.14	*b* = -.07 *SE* = .42 *Z* = -.17 *p* = .86 [-.89; .75] *OR* = 1.09	.34 (.52)	.54 (1.17)	.40 (.67)	*b* = .83 *SE* = .45 *Z* = 1.84 *p* = .07 [-.05; 1.71] *OR* = .44	*b* = .25 *SE* = .45 *Z* = .56 *p* = .58 [-.63; 1.14] *OR* = .78