Addressing Online Political Hostility

Jesper Rasmussen

Addressing Online Political Hostility

PhD Dissertation

Politica

ISBN: 978-87-7335-311-0

Cover: Svend Siune
Print: Fællestrykkeriet, Aarhus University
Layout: Annette Bruun Andersen

Submitted January 31, 2023
The public defense takes place May 4, 2023
Published May 2023

Forlaget Politica
c/o Department of Political Science
Aarhus BSS, Aarhus University
Bartholins Allé 7
DK-8000 Aarhus C
Denmark

Table of Contents

Acknowledgements

Writing a PhD ain't that easy. In fact, it is kind of hard, and for some people it can be a lonely and frustrating journey. Yet for the most part, my PhD journey felt like an enormous privilege. The best (and worst) part of the job is the freedom you have to shape three years of your working life. At the time of writing, I cannot really tell whether the journey was successful or not, but at least I managed to do research which I believe is solid work. In my view, the key to doing a great PhD is having a supportive and competent environment around you that both encourages you to advance as well as challenges you. To the extent that I succeeded in this endeavor, I have to thank a lot of people.

I had an incredible team of supervisors. I am indebted to Lasse Lindekilde for always believing in me and encouraging me to move forward. I am forever grateful for your optimism regarding my academic development, your patience with my absentmindedness, and (perhaps most importantly) teaching me how to catch garfish! It's a pleasure to collaborate with you and I look forward to our next projects! I want to thank Lasse Laustsen for continuously challenging my work and insisting on having me rewrite my manuscripts for the n'th time. Your constructive feedback have undoubtedly improved my work significantly. I also want to thank Michael Bang Petersen for being a stand-in supervisor at the very end of my PhD, collaborating with me as well as hiring me to the ROPH project in the first place. Your ability to address real-world problems through research and communicate your findings clearly to a broad audience is truly inspiring. While writing the summary of this dissertation, I endeavored to emulate your approach, although with varying degrees of success.

A lot of other colleagues at the department have contributed to this dissertation, either directly or indirectly. I am grateful to the PhD group as a whole for a lot of great memories. There are simply too many to name in person (and I would be too embarrassed if I forgot someone), but Andreas Jensen, Matias Engdal, and Ashraf Rachid deserve special praise for their patience and endurance while sharing an office with me or sitting next door. I also had the privilege of being part of an incredible

research environment through the ROPH and STANDBY projects, and I want to thank all current and past colleagues whose valuable feedback contributed significantly to my work. Similarly, I would like to thank my colleagues in the Political Sociology section for being accommodating and helping me settle at the department. The whole TAP-group deserves recognition for always being helpful. In particular, Ruth Ramm, Annette Andersen, Malene Poulsen, Christina Prior, and Kate Thulin have been incredibly supportive, whether it was accounting or language revision. I benefited from excellent research assistance from Nanna Dorthea Olsen, Lea Pradella, and Mathies Jæger Andresen and Christian Noer. Njall Beuschel and Ida Smidt-Jensen deserve special praise for creating a welcoming environment in the staff lounge.

I couldn't have chosen a better place for my research stay than the Social Identity and Morality Lab at New York University in the fall of 2023. I would like to express my sincere gratitude to Jay Van Bavel for hosting me. Jay's enthusiasm, expertise, and generosity were truly invaluable, and I benefited greatly from his guidance. The lab environment is a crucial factor for a successful research stay, and I want to extend my gratitude to Ali Javeed, Claire Robertson, Lina Koppel, Steve Rathje, and Yifei Pei for their friendship and for welcoming me to the lab. You made my stay rewarding both academically and personally. I also want to express my gratitude to Jens Birk, Ken Pound, and Jessie for their kindness.

I am grateful to all who participated in my studies, and especially to those who generously shared their thoughts and experiences with me during the interviews. Without their participation, this research would not be possible and I did my best to adequately represent their stories.

I want to thank all my family and friends for their support and reminding me that there is much more to life than work. Thomas, Tina, Ellen and Carl deserve a special mention for opening their home for us and making our life so much easier during the 3rd year of my PhD when we moved back and forth from New York. Most importantly, thank you Camilla for your love and unwavering support in whatever I do through ups and downs. Thank you for always being there for me and helping me persevere in the face of challenges — both within and outside academia. You are the best thing that has ever happened to me and can't wait to starting the next chapters of our lives together.

Jesper Rasmussen
Aarhus, April 2023

Preface

This report summarizes my PhD dissertation *Addressing Online Political Hostility*. The dissertation was written as the conclusion of my PhD project at the Department of Political Science, Aarhus University. The dissertation consists of this summary report and the research articles that are listed below. The report presents the core contributions of the four articles and discusses the broader implications that cut across the overall theme of addressing online political hostility.

- **Paper A**: Ventilators, Colliders and Megaphones: Pathways To Online Political Hostility On Mainstream Social Media (Submitted)
- **Paper B**: When Do The Public Support Hate Speech Restrictions? Symmetries And Asymmetries Across Partisans In Denmark And The United States (under review)
- **Paper C**: Principled Anti-Egalitarian Values Predict Opposition to Hate Speech Restrictions (under review)
- **Paper D**: Public Health Communication Decreases False Headline Sharing by Boosting Self-Efficacy. Co-authored with Lasse Lindekilde and Michael Bang Petersen (Revise and resubmit at Journal of Experimental Political Science)

Chapter 1
Introduction

In the beginning social media fueled democratic optimism. Being online provided anyone with the opportunity to access information, connect with others and participate in political deliberation (Papacharissi, 2004). Social media was originally expected to advance democracies and topple dictators (Tucker et al., 2017). In 2010, Facebook's CEO Mark Zuckerberg was named "Person of the Year" by Time Magazine for "connecting more than half a billion people and [...] creating a new system of exchanging information and for changing how we live our lives" (Grossman, 2010). What could go wrong? Yet this initial optimism was soon replaced by pessimism, as hostility — particularly in political discussions — – plagued social media platforms (Andresen et al., 2022; Duggan, 2017; Vidgen et al., 2019; Zuleta & Burkal, 2017), leading to concerns that "social media is warping democracy" (Haidt & Rose-Stockwell, 2019).

Policymakers and social media companies face pressures to address behaviors that threaten the democratic potential of social media (Kaye, 2021; Keller, 2019; Suderman, 2018). While social media is often referred to as an open public square that gives voice to democratic forces, behaviors that are corrosive to democratic norms are frequent. Public deliberation relies on norms of free and open debate in which informed decisions are based on accurate information. Yet some behaviors on social media undermine democratic norms by being actively hostile towards those same norms — I refer to these behaviors as online political hostility.[1] Such behaviors involve the sharing of misinformation that undermines decision-making based on rational and informed public deliberation, or the use of hate speech that threatens free and equal participation in discussions. The common denominator is that online political hostility poses a challenge for deliberative norms on social media and needs to be addressed.

A large group of explanations for online political hostility suggest that these forms of malevolent behavior are shaped by non-political fac-

[1] I elaborate on this definition in Chapter 2.

tors such as people's personalities or contextual factors on social media that trigger or incentivize hostility. One of the most predominant narratives emphasizes how internet "trolls" — i.e., sinister individuals who disrupt online discussions for the amusement — are ruining the online sphere (Stein, 2016). According to this account, hostility is motivated by aggressive or anti-social personalities that undermine online interactions "for fun" (Buckels et al., 2014; Eberwein, 2019; Erjavec & Kovačič, 2012). In a similar vein, other accounts suggest that people are hostile because of the affordances of social media (Cheng et al., 2017; Wolchover, 2012). People may be inattentive to accuracy (Pennycook et al., 2021) or unaware of the consequences of their behavior , while anonymity and the absence of non-verbal cues change people's behavior for the worse (Suler, 2004). At their core, these explanations suggest online political hostility can be addressed by nudges or platform design, because online political hostility is thought to result from flaws of the interplay between social media and human psychology, rather than to be deliberate political action.

Yet there is good reason to believe that being hostile in political discussions on social media is a deliberate political act. People who are active on social media are more politically extreme and polarized, which in turn motivates engaging in online political hostility (J. W. Kim et al., 2021; Osmundsen et al., 2021; Wojcieszak et al., 2022). Rising levels of polarization (Iyengar et al., 2019; Iyengar & Westwood, 2015) intensify conflict and animosity on social media (Ruggeri et al., 2021; Van Bavel, Rathje, et al., 2021). Furthermore, the sharing of hate speech and misinformation spikes around political conflicts in the real world such as elections or protests (Grinberg et al., 2019; T. Kim, 2022; Rasmussen & Petersen, 2022; Siegel et al., 2019). These explanations suggest that politics is at the root of the behaviors on social media that undermine democratic norms, yet politics is often neglected in interventions that seek to address online political hostility. In other words, some of the most widespread interventions assume that online political hostility is largely apolitical and thus can be "corrected" if only people are nudged in the right direction.

The point of departure for this dissertation is to assess this assumption and, on this basis, provide an answer to *how online political hostility can be addressed* by interdiction through regulation or mitigated by building competences among the audience. I advance the arguments of this dissertation through four self-contained research articles. First, Paper A examines one of the most prevalent assumptions of online political hostility: that psychological flaws or contextual features of social media

shape online political hostility (Buckels et al., 2014; Cheng et al., 2017; Pennycook et al., 2021). I argue that these lines of research neglect the deliberate political meaning that people attribute to their own behavior, and assess this proposition through 25 interviews with people who have engaged in online political hostility in Denmark. Through these interviews, I outline three distinct pathways to online political hostility: Ventilators engage in hostility to seek relief from their political frustrations; colliders engage in heated political discussions online, which spawns collisions; and megaphones use social media to persuade other people and gain influence, occasionally through hostile measures. These pathways underline that while there are many avenues to online political hostility, it is often motivated by political beliefs, frustrations and opinions. In other words, online political hostility is deliberate political activism, rather than random disruptive or apolitical behavior. Because hostility reflects real political frustrations, this behavior is harder to change and requires addressing its root causes, including lack of trust, inequality and feelings of marginalization. In the long term, policymakers and social media platforms should strive to increase trust and transparency and mitigate the factors that foster online political hostility — not least because these frustrations are encountered in people's daily lives. However, the structural changes that are needed to address the root causes of online political hostility as suggested in Paper A are not likely to happen in the short term. In the meantime — while we wait for structural change — what can policymakers and social media platforms do in the short term to address online political hostility? If online political hostility should be addressed in the short term, one potential avenue is to change the focus from the perpetrators of online political hostility to its audience and victims. In this dissertation, I propose two tools that either protect social media users from online political hostility through regulation or empower them by building certain competences.

Regulation of online content is associated with concerns about free speech, and one of the primary concerns is that people cannot agree what hate is and want to censor their political opponents. Yet in Paper B I argue that people do consistently want to restrict extreme speech, regardless of political affiliation or of whom it targets. Contrary to popular accounts, people are not more likely to censor political opposition. In other words, this suggests that there is more consensus on when online political hostility should be restricted than previously thought. Online political hostility most likely can be addressed by regulating based on the severity of the content, because the public agrees that severity is the key criterion for regulating online political hostility. Yet at the same time, Paper B shows

that people on the political right do have more reservations in terms of regulating online political hostility. What shapes these varying degrees of opposition across the political spectrum? In Paper C, I provide an empirical assessment of the two primary accounts of these differences. One account suggests that people oppose regulating online political hostility because it is a tool to derogate minority or historically oppressed groups (Bilewicz et al., 2017). At its core, this account assumes that opposition to regulating online political hostility is based on group-based dominance motivations (Federico & Sidanius, 2002; Sidanius et al., 1996). Yet another account suggests that people oppose regulating online political hostility because of "principled conservatism" (Sniderman & Carmines, 1997; Sniderman et al., 1991; Sniderman & Piazza, 1993; Sniderman & Tetlock, 1993) in the form of anti-egalitarian values such as limited government and a free market of ideas. I demonstrate empirical support for the latter. Thus attitudes towards regulating online political hostility are shaped by principled political values, rather than prejudice and dominance orientations per se.

Another strategy for addressing online political hostility is empowering citizens by fostering competences that mitigate the adverse effects of other people's engagement in online political hostility. In Paper D, I argue that forms of communication designed to build competences among citizens reduce the sharing of online political hostility and make citizens feel more efficacious. In other words, providing clear, elaborate and concrete advice can make citizens share less political hostility online. I advance this argument by evaluating three interventions that aimed at reducing misinformation about COVID-19. I find that equipping people with concrete, specific and actionable advice boosts their feelings of competence and in turn decreases their misinformation sharing. Thus, interventions that build competences reduce the spread of online political hostility when democracies face crises (such as the onset of pandemics), at least in the short term.

In summary, the overall argument of the dissertation is that online political hostility is a form of deliberate political activism. As these motivations are stable and steadfast and require long-term policy changes, I propose that online political hostility can be addressed by shifting the focus from the perpetrators to the audience. Specifically, I propose two tools to address online political hostility through regulation and empowerment. I show that while principled political values shape opposition to regulating online political hostility, people do agree that severity is the key criterion for regulation. Furthermore, citizens can be empowered to

share less online political hostility when they are equipped with concrete tools and advice.

The dissertation proceeds as follows. In Chapter 2, I start by conceptualizing online political hostility and how it serves as an umbrella term for behaviors on social media that undermine democratic norms of public deliberation, including hate speech and misinformation. I then turn to providing a short review of the political and non-political accounts of why people engage in online political hostility. Finally, I propose that online political hostility can be addressed by shifting the focus from the perpetrators to the audience through regulation and empowering interventions. In Chapter 3, I start by providing an overview of the research designs in this dissertation. I conclude this chapter by discussing the strengths and limitations of the dissertation in relation to standard research criteria. Then in Chapter 4, I outline the core findings of this dissertation. The chapter starts by examining the assumptions regarding why people engage in online political hostility and then proceeds to present empirical results on how it can be addressed through regulation and empowerment. Finally, in Chapter 5 I discuss the dissertation's findings in relation to existing knowledge. In short, I argue that there is no quick fix to online political hostility. Hostility on social media is a reflection of deep-rooted dispositions, political motivations and frustrations. Yet at some points—particularly during times of crisis—interventions focused on empowerment may be a viable tool in mitigating online political hostility. We should be less concerned that people may seek to censor ideas and groups that they don't like. Rather, people's willingness to censor is characterized by a consensus that extreme speech should be restricted. I do find, however, that people disagree upon the threshold of extremity. I conclude by outlining potential avenues for further research, including how activating bystanders may foster and preserve democratic norms.

Chapter 2
Theory

This chapter outlines the theoretical framework of the dissertation. I start by defining online political hostility, which serves as an umbrella term for online behavior that undermines democratic norms, and which includes hate speech and misinformation. I then turn to reviewing why people engage in online political hostility and group these explanations into two categories. The first are those that emphasize psychological flaws and contextual features of social media as explanations for engaging in hostility online, and the second are those that emphasize politics. Finally, I outline how online political hostility can be addressed by shifting the focus to the audience instead of the perpetrators through strategies of empowerment and regulation.

2.1 Conceptualization of Online Political Hostility

Public deliberation is a key element of a well-functioning democracy. Deliberation entails debate and discussion "aimed at producing reasonable, well-informed opinions in which participants are willing to revise preferences in light of discussion, new information, and claims made by fellow participants" (Chambers, 2003). Yet deliberation sets a range of requirements. Its content should contribute to well-informed opinions and there should be broad access to participation in the debate, and those participants should be willing to revise their previously held opinions in the face of new evidence. Such deliberation relies on free and open debates in which accurate information is shared to inform decisions.

When citizens participate in public discussions on social media, they sometimes do so in a way that undermines deliberative norms (Quandt, 2018). Some might — intentionally or unintentionally — disseminate false information that leads the public astray, undermining the pursuit of well-informed opinions. Others might use their freedom of expression

to damage others' reputations through slander. In the most severe cases, people use intimidation or threats to prevent their political opponents from participating in political debates, eroding norms of free and equal access to debate.

These types of behavior lie at the core of this dissertation. I use online political hostility as an umbrella term for behaviors on social media that undermine democratic norms of public deliberation. Online political hostility encapsulates the forms of "public-level incivility" that are "violations of political process and deliberative norms" (Muddiman, 2017). Violating democratic norms may range from degrading others or spreading misinformation to more extreme types of hostility such as hate speech in the form of sexist, racial or religious slurs, dehumanization or threats (Rossini, 2019). Importantly, online political hostility is distinct from impoliteness, negativity or conflict in and of itself, although "violat[ing] norms of politeness for a given culture"(Mutz, 2015, p. 6) may also influence deliberation (Massaro & Stryker, 2012). Rather, online political hostility concerns violations that "threaten a collective founded on democratic norms" (Papacharissi, 2004, p. 271) and thus hinder "public discussions and carefully weighing a comprehensive set of ideas" (Muddiman, 2017). Social media provides a venue for people to have political discussions online, yet online political hostility undermines some people's free and equal access to public discussions, genuine exchange and rational arguments (Stromer-Galley & Wichowski, 2011).

The papers of this dissertation primarily focus on two sub-categories of online political hostility: hate speech and misinformation. The definition of hate speech — both online and offline — is an unsettled question across public, academic, legal and policy debates (Siegel, 2020). As noted by Gagliardorne and colleagues, "hate speech continues largely to be used in everyday discourse as a generic term, mixing concrete threats to individuals' and groups' security with cases in which people may be simply venting their anger against authority" (Gagliardone 2016). Given this kind of ambiguity, it is perhaps not surprising that according to a public poll of Americans, 82 percent believe "that it would be difficult to ban hate speech because people can't agree what speech is hateful and offensive" (Cato, 2017). The United Nations defines hate speech as "[A]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" (ICCPR, 1966). Similarly, within the social sciences, hate speech is often referred to as "bias-motivated, hostile, and malicious language targeted at a person or group because of their actual or perceived innate characteristics" (Cohen-Almagor, 2011). These definitions capture two of the key components of most hate speech

definitions: the severity of the content and the target of the statement. Thus, in its broadest sense hate speech portrays a group negatively and may range from offensive speech to incitements of violence (Mchangama et al., 2020). This entails intimidation of other people by derogating, dehumanizing or even threatening them. Groups are defined by protected characteristics such as ethnicity, sexual orientation or disability. While there is considerable variation in whether and which groups are protected, hate speech undermines democratic norms by threatening the dignity, liberty and equality of others, which in turn limits people's free and equal access to participation in political discussions (Gagliardone et al., 2016; Sellars, 2016).

Misinformation is also a disputed concept, often used interchangeably with concepts such as disinformation or fake news. Various definitions emphasize different factors, including whether the information is fabricated (Lazer et al., 2018) and the extent to which the information is intentionally and verifiably false (Allcott & Gentzkow, 2017). I apply a broad understanding of misinformation as "false or misleading messages spread under the guise of informative content [...] constituting a claim that contradicts or distorts common understandings of verifiable facts" (Guess & Lyons, 2020, p. 10). As such, misinformation is false by definition, and undermines the democratic norm of informed public deliberation based on accurate information because it inhibits rational discussion based on facts.

While hate speech and misinformation are distinct concepts, they sometimes share considerable overlap. For instance, hostile rumors share features of both misinformation and hate speech, as they "portray politicians and political groups negatively and possess low evidential value," and at the same time seek to "incite hostility toward a specific target" (Petersen et al., 2018). Thus, hostile rumors share commonalities with conspiracy theories, negative campaigns and misinformation as well as hate speech. In sum, hate speech and misinformation are two forms of online political hostility—an umbrella term for behaviors on social media that undermine deliberative democratic norms. Hate speech and misinformation are characterized by definitional ambiguity and are related to forms of incivility. Yet online political hostility is distinct from interpersonal incivility, because it undermines democratic norms including equality of participation, reciprocity and rational argumentation.

Online political hostility emerges in the context of political discussions on social media or when hostility is directed against politicians and people who participate in public debates on social media. Politics entails activities that distribute resources and status and have implications for all

citizens, in contrast to private affairs. As people have different interests in the distribution of resources, conflict is an inherent feature of politics (e.g. Easton, 1965). While people can influence political decisions through public deliberation, access is key. If public deliberation is hostile — in the sense of being antagonistic towards democratic norms — people may withdraw from participating in political deliberation. Evidence suggests that political talk is indeed more hostile in the online sphere (Andresen et al., 2022; Bor & Petersen, 2021), and studies suggest that many people avoid both talking politics and exposing themselves to opposing partisan news (Mukerjee & Yang, 2020). Thus, the "political" element in online political hostility refers to politics as an arena rather than specific content or opinions, and is not limited to negativity, as politics is inherently conflictual in nature.

The factors that motivate online political hostility can be both political and non-political. Hostility can be a deliberate strategy to articulate political viewpoints and raise attention in public political discussions, but it can also be a consequence of inattention, failure to control emotions or sinister sadistic motivations. Thus, online political hostility is not defined by its motivations, but by its undermining of democratic norms in political discussions in the public sphere on social media. Yet the motivations that underlie this behavior are not trivial — rather, they are crucial in terms of addressing online political hostility. In the following section, I provide a short review of the political and non-political accounts often used to explain online political hostility.

2.2 Why do people engage in online political hostility?

Mainstream social media platforms are the key arena for political deliberation online. For instance, 75% of Danes have a Facebook account. Facebook is also the place where most Danes are exposed to online political hostility, particularly when they engage in political discussions online (e.g Andresen et al., 2022; Duggan, 2017; Vidgen et al., 2019). Thus, mainstream social media platforms are the key arena for political deliberation, but also where people are most likely to encounter online political hostility. This in turn makes some people reluctant to participate in political discussions online, which constitutes a problem for political deliberation. Why, though, do people engage in political hostility on social media?

While extensive research has documented how various forms of hostility prevail in closed, extreme networks (Bliuc et al., 2018), less is known about the producers of online political hostility on mainstream platforms (Siegel, 2020). In this section, I provide a brief review of the predominant explanations for why people engage in online political hostility. I group these explanations into two categories: those arguing that online political hostility is a motivated political act and those arguing that it rather reflects personality traits or is the consequence of certain features of social media. The latter group of explanations—which I refer to as non-political — paint a bleak picture of people who engage in online political hostility as disagreeable individuals who derail political discussions because they want to humiliate others, fail to control their emotions or strive for status through dominance. Yet few studies have sought to examine people's own accounts of why they engage in online political hostility, making it hard to develop interventions that resonate with them. I conclude by proposing that systematic qualitative research designs are needed in order to understand these motivations.

2.2.1 Non-political accounts

A range of explanations of why people engage in online political hostility assume that the hostility is not motivated by politics, but reflects "dark" personality traits, psychological flaws and the features of social media. Similar to arguments that politics is not the reason why people engage in political discussions in general (Hersh, 2017), these explanations highlight that people do not seek to undermine democratic norms on social media because of politics. Rather, these lines of research highlight that antisocial personality traits, sadistic motivations to humiliate others or the features of social media shape hostility in different ways. In the following, I review some of the main claims within these literatures.

As a starting point, it is a "reasonable position to take that the structure of people's online social networks and the types of communication they engage in via electronic media are relatively similar to their real life counterparts. After all, it is the same person engaging in these behaviors in both scenarios" (Crosier et al., 2012). Indeed, evidence suggest that people's engagement in online and offline political hostility (Bor & Petersen, 2021) as well as related behaviors such as bullying (Kowalski et al., 2014; Kowalski et al., 2019) are highly correlated. Yet, as noted by Crosier and colleagues (Crosier et al., 2012), "different situations and contexts encourage the differential expression of steadfast traits." In turn, we should expect the interplay of individual dispositions in combination

with specific situational and system-level factors to influence engagement in online political hostility.

Online political hostility is shaped by a range of stable as well as malleable characteristics at the individual level. Regarding the more stable characteristics, a range of studies have examined the relationship between personality traits and forms of online political hostility. People who are less agreeable engage more in online political hostility. For instance, (dis)agreeableness is related to bullying in online environments (Van Geel et al., 2017). A recent meta-analysis of dark personality traits and antisocial online behaviors suggested that "psychopathy was the trait most strongly and consistently correlated with the majority of the explored antisocial online behaviors, followed by Machiavellianism and everyday sadism" (Moor & Anderson, 2019). Trolling is one of the disruptive behaviors that has received the most attention in the past decade. Trolls are often described as goblin-like individuals who disrupt, derail and ruin genuine discussions without a specific cause, values or beliefs (Stein, 2016). Rather, trolls are sadistic and motivated by the thrill and fun of humiliating others and the cascade of online dynamics this prompts (Buckels et al., 2014). Social media provides amusement, and trolls are said to enjoy the entertainment value of inflicting harm and anger on others (Eberwein, 2019; Erjavec & Kovačič, 2012). In line with this view, Van Geel et al. (2017) suggest that people who engage in cyberbullying score lower on agreeableness and higher on everyday sadism.

Another individual difference that is related to forms of online political hostility is social dominance orientation—that is, people's preference for maintaining and enhancing group-based hierarchies. Social dominance orientation is associated with prejudice, racism, sexism and acceptance of hate speech (Ho et al., 2015; Pratto, 1994). Engaging in online political hostility on social media can thus work as a tool to enhance group-based dominance hierarchies through public aggression (Sidanius & Pratto, 2001). Minority and historically oppressed groups are indeed the targets of most hostility online (Andresen et al., 2022).

Finally, a range of explanations highlight how the design and incentive structures of social media fuel hostility in political discussions. These suggest that incidents of hostility might be "accidents" triggered by situational features that can make even ordinary citizens become trolls (Cheng et al., 2017). On social media, people may be distracted (Pennycook et al., 2021) or unaware of the consequences of their behavior (Suler, 2004), while anonymity and absence of non-verbal cues from the offline

world may change people's behavior for the worse (Rowe, 2015; Stein, 2016; Wolchover, 2012).

Overall, these explanations suggest that online political hostility is motivated by factors that are not political, but rather reflect personality traits and contextual and situational features of social media environments.

2.2.2 Political accounts

Social media initially prompted democratic hope. Yet the optimism soon faded as it became clear that online political discussions turned out to be more hostile than expected. Indeed, in the previous section, I outlined how non-political factors lead people to online political hostility in discussions on social media. Yet research suggests that hostility is particularly likely to spawn in political discussions. In this section, I review the literature on how politics is related to being hostile in political discussions on social media.

A large group of explanations suggest that hostility on social media is driven by political motivations, events and polarization. An extensive literature suggests that affective polarization — the tendency to view opposing partisans negatively and co-partisans positively — is on the rise and fuels partisan animosity (Iyengar et al., 2019; Iyengar & Westwood, 2015). Particularly in the United States, partisans hold exaggerated perceptions of the level of partisan animosity that one's political outgroup feels about one's ingroup (Ruggeri et al., 2021) or are subject to false polarization — people's tendency to overestimate the degree of polarization between groups (Levendusky & Malhotra, 2016). Partisans exaggerate their political outgroup's opposition to democratic norms (Pasek et al., 2022), the extent to which they hold prejudice towards and dehumanize one's ingroup (Cassese, 2021; Kteily et al., 2016; Martherus et al., 2021; Moore-Berg et al., 2020; Pacilli et al., 2016) and whether their outgroup is willing to obstruct their ingroup for political gain (Mernyk et al., 2022). In turn, affective polarization and (mis)perceptions may increase conflicts and spawn online political hostility in various ways. One study suggests that partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter, and the sharing of misinformation is driven by the same psychological motivations as sharing partisan news from traditional and credible news sources. They conclude that "[...] individuals who report hating their political opponents are the most likely to share political fake news and selectively share content that is useful for derogating these opponents." Thus, when polariza-

tion is high, we should expect that people are more likely to share content — regardless of its veracity — for political gain (Osmundsen et al., 2021). Indeed, political factors seem to be heavily related to online engagement, including its hostile forms. One study suggested that frequent commenters were more interested in politics, held more polarized opinions and in turn used more toxic language when commenting (J. W. Kim et al., 2021). Furthermore, content with negative moral emotions (Brady et al., 2017; Robertson et al., Forthcoming) or that attacks political opponents (Rathje et al., 2021; Yu et al., 2021) increases engagement.

Aside from motivations related to classic ideological drivers, online political hostility can be motivated by disruptive political sentiments. One line of research suggests that the interplay of social marginalization and status-oriented personalities predicts sharing of hostile political rumors (Petersen et al., 2018). Online political hostility may serve as a strategy to attain status for people who nurse grievances in terms of their position in society. In other words, social and political frustrations shape online hostility. Furthermore, research suggests that disruptive political goals are related to belief in and sharing of conspiracy theories and fake news. One study shows that "anti-establishment" orientations—that is, orthogonal to the classic left-right dimension of public opinion—are related to the acceptance of political violence, time spent on extremist social media platforms and belief in misinformation and conspiracy theories (Uscinski et al., 2021). In other words, desires to disrupt the political system are related to online political hostility.

Finally, political events in the offline world are a stable predictor of online political hostility. Studies suggest that various forms of online political hostility, including hate speech and misinformation, spike around elections and then quickly return to their "normal" levels (T. Kim, 2022; Rasmussen & Petersen, 2022; Siegel et al., 2019). For instance, during the final weeks of the 2016 US presidential election, the prevalence of fake news suddenly increased before returning to its previous level immediately after the election (Grinberg et al., 2019). Various forms of political crises including the COVID-19 pandemic, protests, riots and wars are also associated with spikes in online political hostility (Rasmussen & Petersen, 2022). Historically, hostile rumors serve key functions in escalating intergroup conflict (Horowitz, 2001) and may serve to signal group affiliation, mobilize group members and coordinate action (Petersen et al., 2020). Recent studies document that online political hostility spiked during heavy influxes of refugees, Black Lives Matter protests and the Capitol Hill insurrection (Hangartner et al., 2019; Rasmussen & Petersen, 2022). These findings suggest that online political hostil-

ity does not occur in a vacuum, but is heavily influenced by events in the offline world. Furthermore, political events shape when and where online political hostility is likely to flourish. In sum, these accounts highlight how politics influences engagement in online political hostility, both through events in the offline world and through political conflicts at the group level.

Overall, the literature suggests a range of political and non-political explanations of why people engage in online political hostility. In general they paint a bleak picture of those who engage in hostility, making it hard to intervene in ways that resonate with them. Yet very little is known about people's own accounts of engaging in political hostility on mainstream social media platforms (Siegel, 2020, p. 56, 61–64). In other words, do they think of their behavior as political participation? Some of the few studies seeking to understand the meaning that people ascribe to their own behavior highlight the instrumental value of engaging in online political hostility by interviewing some of its perpetrators on news website comments sections. People employ aggressive strategies to further their political causes. Some rationalize that hostile language is a tool to make people recognize the truth (Eberwein, 2019; Erjavec & Kovačič, 2012), and these people often consider themselves to be particularly knowledgeable (Fangen & Holter, 2019). In other words, engaging in online political hostility is perceived as a way to further a political cause on news website comments sections. These studies mark an important shift in understanding people's rationales and justifications for online political hostility. In order to address online political hostility, it is indeed necessary to understand why people engage in online political hostility from their own perspectives.

Despite these initial advances in the literature, state-of-the-art interventions countering online political hostility still predominantly emphasize non-political factors. Interventions often assume that the source of the hostility is psychological flaws, rather than genuine political frustrations. Yet as noted by Siegel (2020, p. 56) on the responses to various forms of online political hostility, "governments worldwide are passing regulation and pressuring social media companies to implement policies to stop the spread [yet] these calls for action have rarely been motivated by comprehensive empirical evidence [and] researchers have only recently begun to examine the efficacy of approaches to countering online hate, and our understanding of the collateral costs of these interventions is especially limited." In order to address political hostility on social media, there is a need to test the assumptions through systematic quali-

tative studies that can account for why people engage in online political hostility.

If politics lies at the core of engaging in online political hostility, people who do so are using the internet and social media in exactly the way that was originally envisioned: as a public square where politics can be discussed. Thus, the interventions addressing online political hostility by correcting its perpetrators likely won't work, because people are expressing their political beliefs and opinions. In other words, online political hostility might not be the product of psychological flaws. Rather, it may be a deliberate form of political participation. If so, interventions should change their perspective, moving from addressing the producers of online political hostility to protecting or empowering its victims and audience. In the next section, I start by reviewing the assumptions of some of the state-of-the-art interventions addressing online political hostility and outline ways of empowering and protecting the public.

2.3 How can online political hostility be addressed?

Addressing online political hostility requires a proper understanding of why people engage in it in the first place. In the previous sections, I outlined two groups of explanations highlighting political and non-political factors for engaging in online political hostility and highlighted that state-of-the-art interventions addressing online political hostility predominantly emphasize non-political factors. Thus, interventions often assume that hostility has its origins in psychological flaws, rather than genuine political frustrations. For instance, fact-checking assumes people fall for misinformation because they don't know any better (Carey et al., 2022). Accuracy nudges assume that people "forget" that they are intrinsically motivated to share accurate news (Pennycook et al., 2021). Digital literacy interventions are based on people lacking cognitive competences to engage on social media (Guess et al., 2020). And finally, interventions that emphasize empathetic norms assume that people don't know that sharing online political hostility is hurtful (Hangartner et al., 2021; Munger, 2016, 2020; Siegel & Badaan, 2020). While these interventions may be effective under certain circumstances, they are largely based on the assumption that the motivations for engaging in online political hostility are nonpolitical. If the motivations are political, however, then addressing the root cause of online political hostility requires structural and long-term policy change. In other words, addressing the motivations for

online political hostility is hard in the short-term. Instead, social media platforms and policymakers can opt for strategies that protect online political hostility's audience.

In this dissertation, I examine an interdiction and a mitigation strategy against online political hostility addressing it through regulation and empowerment respectively. These strategies take as their starting point that while it is hard to change people's motivation, witnesses can be protected from the negative consequences of online political hostility through interdiction or mitigation strategies. In other words, if the motivations of the perpetrators cannot be changed in the short term, another viable strategy is to protect or empower the audience. In this view, regulating content reduces exposure to online political hostility, while providing tools to the audience empowers people with respect to how they can react when online political hostility does emerge. In the following, I address these two in turn.

2.3.1 Regulating online political hostility

Instead of changing the motivations of perpetrators, interdiction strategies seek to regulate online political hostility. Such measures are not without controversy, as freedom of expression is a cornerstone in democracies and is a fundamental human right. The right to seek, receive and impart ideas — even those that are disagreeable — and information of all kinds regardless of borders reflects classic liberal values (ICCPR, 1966; Mill, 1966). Yet democracies face a dilemma regarding, on the one hand, protecting individuals' rights to express their ideas and, on the other hand, protecting the rights of others to dignity and respect (e.g. OHCHR, 2012).

Regulating social media is not easy. Yet, as people do frequently experience political hostility on social media (Andresen et al., 2022), social media platforms are increasingly faced with pressures to regulate the sites. At the same time, concerns about the erosion of democratic norms and civil liberties are rising (Bartels, 2020; Carey et al., 2019). In the past decade, there were no liberties that "deteriorated as much as those related to freedom of expression and media freedom," according to the Economist's Democracy Index. In western liberal democracies, the decline in free speech is driven by laws that regulate misinformation and hate speech (EIU, 2020). In turn, these developments spawned concerns about the censorship of political opposition, most prominently illustrated by the suspension of former US president Donald Trump from Twitter following the attack on Capitol Hill. Most Americans indeed think that

social media platforms censor political viewpoints (Vogels et al., 2020) and believe that people cannot agree on whether and how to regulate online political hostility (Cato, 2017; Dunn, 2019).

One of the key challenges for interdiction is that people oppose regulating online political hostility. One line of research suggests that this opposition is based on principled political values deriving from conservatism. This form of *principled conservatism* emphasizes preferences for limited government and a free market of ideas (Sniderman & Carmines, 1997; Sniderman et al., 1991; Sniderman & Piazza, 1993; Sniderman et al., 1989). In contrast to this perspective, the *group-based dominance* perspective suggests that people oppose regulating online political hostility because it is a useful tool for dominating minority and historically oppressed groups. In other words, the opposition stems from prejudice rather than ideology (Sidanius et al., 1994; Sidanius & Pratto, 2001; Sidanius et al., 1996). Indeed, previous studies have suggested that group-based dominance is related to racism, sexism, prejudice (Kunst et al., 2017) and hostility toward minority and historically oppressed groups (Bilewicz et al., 2017; Costello & Hodson, 2011; Esses et al., 2008; Kteily et al., 2015). In sum, these two perspectives debate whether opposition to regulating online political hostility is shaped by "racism or is based instead on a principled objection to the nature" of the regulations (Feldman & Huddy, 2005).

Other lines of research highlight how opposition to regulation is shaped by contextual features. A classic finding suggests that people are intolerant towards ideas and groups they dislike (Stouffer, 1955). People endorse abstract ideals of free speech, yet when they are asked in concrete settings, they are reluctant to extend civil liberties to disliked groups (Marcus et al., 1995; Sullivan et al., 1982). Especially when people perceive threats towards their own group or society, they are likely to exhibit intolerance (Gibson, 2011, p. 418). This suggests that the target of speech matters for intolerance, which in turn may have consequences for people's willingness to censor opposition on social media.

Recent research indeed suggests that some people are ready to trade democracy for partisanship (Frederiksen, 2022; Graham & Svolik, 2020; Simonovits et al., 2022; Svolik, 2018), which eventually would alter the rules of the game in liberal democracies. Recent evidence suggests that partisanship shapes perceptions of hostile rhetoric (Muddiman, 2017; Mutz, 2015; Stevens et al., 2015) and censorship of political opponents (Amira et al., 2021; Ashokkumar et al., 2020; Lelkes & Westwood, 2016). Furthermore, some partisans tend to dehumanize members of the opposing party (Martherus et al., 2021; Moore-Berg et al., 2020), and exagger-

ate their political opponents' willingness to engage in political violence (Kalmoe & Mason, 2022). Based on this, the challenge for interdiction strategies is to find common ground when regulating online political hostility, as the existing literature suggests that people have different rationales for regulating and may be biased in political contexts.

2.3.2 Empowering the audience

There are good reasons to believe that social media platforms and politicians cannot solve the emergence of online political hostility in the short term. Instead, mitigation strategies targeting the audience that faces online political hostility may prove effective. A range of interventions target the audience when they are exposed to various forms of online political hostility (see Van Bavel, Harris, et al., 2021, for an overview). In this dissertation, I focus on those interventions that provide tools, skills or competences to the audience, even though a range of interventions rely on people's intrinsic accuracy motivations by reminding people to "think before they post" (e.g. Pennycook et al., 2021). In short, I focus on interventions that empower people. Empowering the audience is important because a small number of people account for most of the engagement in online political hostility (e.g. Grinberg et al., 2019; Osmundsen et al., 2021). For instance, Grinberg et al. (2019) show that 0.1% of users were accountable for 80% of the sharing of fake news during the 2016 American presidential election. This suggests that a majority of the people who encounter forms of online political hostility are the audience, not the perpetrators.

At their core, interventions that empower the audience equip them with competences in terms of how they should respond when they face online political hostility. Some interventions are based on inoculation theory and follow the biomedical analogy that we can inoculate people against engaging in hostility (McGuire & Papageorgis, 1962). These interventions seek to preemptively forewarn and expose people to weakened doses of misinformation alongside strong refutations, which in turn cultivates cognitive resistance against future misinformation (Roozenbeek et al., 2022; Van Der Linden, 2022). These kinds of interventions have proven effective both in the short and long term (Maertens et al., 2021).

Other efforts focus more explicitly on providing competences through tips and instructions as part of digital literacy and bystander interventions (Guess et al., 2020; Hertwig & Grüne-Yanoff, 2017; Lee, 2018; Rudnicki et al., 2022; Sheeran et al., 2007; Sheeran & Orbell, 2000). Feelings

of competence are important in terms of responding effectively to threats and risks such as online political hostility (Maddux & Rogers, 1983), and are related to protective behaviors without inducing fear (Jørgensen et al., 2021). A long line of research shows that people respond effectively when they are made aware of a threat, told how to respond and are assured that the response is efficient (Rippetoe & Rogers, 1987; Rogers, 1975). The common denominator of these interventions is that they empower the audience against online political hostility by using tools and competences as a mitigation strategy.

Chapter 3
Methods

In this chapter, I present the research designs I employed to assess how online political hostility can be addressed, including the most central methodological choices of the dissertation. The structure of the chapter starts out with an overall description of the research designs in Table 3.1. I then turn to describing the core elements of the papers in the dissertation, highlighting and discussing key choices of the research designs in light of relevant research criteria.

3.1 Overview of studies

Table 3.1 provides an overview of the data collection for the four papers and their key methodological contributions to the dissertation. I employed a range of diverse methodological approaches to further our knowledge about how online political hostility can be addressed. For instance, in Paper A I utilized interviews to understand the meaning that people ascribe to their own behavior, while Paper B utilizes experiments to disentangle the causal effects of attributes shaping puiblic opinion on regulating online poltical hostility. While these approaches build on distinct research traditions, they all contribute to building a mosaic of knowledge that helps answer *how online political hostility can be addressed* from different perspectives, which I elaborate on in the following.

3.2 Understanding online political hostility

To address online political hostility, it is important to understand why people engage in it. Yet many of the interventions and policies designed to counter online political hostility assume — either implicitly

Table 3.1: Overview of data collection and key research criteria of this dissertation

Key criterion	Description	Design
Paper A Measurement validity and ecological validity	To assess why people engage in online political hostility I conducted 25 interviews with individuals who had previously been hostile in political discussions on mainstream social media platforms in Denmark	Semi-structured interviews
Paper B Internal validity and external validity	To disentangle which features of online political hostility that shape support for regulation two conjoint experiments were embedded in surveys fielded to nationally representative samples in Denmark (n = 1518) and the United States (n = 1535) with a full sample size of 3053	Conjoint experiments
Paper C Measurement validity and external validity	To examine whether opposition to regulating online political hostility is shaped by anti-egalitarian values (SDO-E) or dominance (SDO-D) I relied on cross sectional data and conjoint experiments embedded in surveys fielded to nationally representative samples in Denmark (n = 1518) and the United States (n = 1535) with a full sample size of 3053	Cross section analysis and conjoint experiments
Paper D Internal validity and ecological validity	To assess the effectiveness of the interventions a nationally representative two-wave panel with embedded survey experiment was employed in Denmark ($n_{wave\ 1}$ = 2541, $n_{wave\ 2}$ = 2232) and to assess whether it boosted their feelings of competence another nationally representative sample was collected in Denmark (n = 2012)	Survey experiments

Note: Data for Paper B and Paper C were collected in the same survey.

or explicitly — that the hostility is not a political act, even when it occurs in political discussions. Rather, most focus on addressing a lack of cognitive resources (Guess et al., 2020; Hertwig & Grüne-Yanoff, 2017; Lee, 2018; Pennycook et al., 2021; Roozenbeek et al., 2022; Rudnicki et al., 2022; Sheeran et al., 2007; Sheeran & Orbell, 2000; Van Bavel, Harris, et al., 2021; Van Der Linden, 2022). Thus, in the first data infrastructure, I start by examining the assumption that it is factors other than politics that make people hostile in political discussions on social media.

The vast majority of research on various forms of online political hostility utilize variance-based designs. These studies aim at generating inferences about factors associated with online political hostility (e.g. Buckels et al., 2014; J. W. Kim et al., 2021; Moor & Anderson, 2019). Some of the most advanced designs couple various forms of text analysis with panel or registry data to generate and provide causal claims (e.g. Guess et al., 2021). Yet one of the key drawbacks of the variance-based approaches is that they cannot account for the meaning that the producers of hostility in political discussions ascribe to their own behavior. In other words, variance-based designs showing that certain personality traits are associated with online political hostility may suggest that people who have more sadistic personalities engage more in hostility, but they tell little about people's own account for their behavior. Essentially, this is a measurement problem which I address in Paper A because we do not fully understand the motivations that underlie online political hostility, we might mistakenly infer that people who act or look like "trolls", are not motivated by politics. This is important, as it has consequences for whether and how online political hostility can be addressed.

To assess the motivations for engaging in online political hostility, in Paper A I conducted 25 in-depth interviews between September 2021 and June 2022 with people who engaged in online political hostility in Danish Facebook or Twitter comments sections. Previous studies have examined producers of hostility in the comments sections of news websites (Eberwein, 2019; Erjavec & Kovačič, 2012; Fangen & Holter, 2019; Faulkner & Bliuc, 2016; Ihlebæk & Holter, 2021), yet I focus on mainstream social media platforms, both because they are the primary venues for public deliberation and sources of information and because they are where most people experience hostility — particularly in political discussions, which leads some people to opt out of these discussions altogether (Andresen et al., 2022; Zuleta & Burkal, 2017).

In other words, this study examines why people engage in hostility in the places where other people are most likely to experience it.

Interviewees were recruited[1] based on their participation in political discussions on Facebook or Twitter in which they posted a hostile comment or reply to public posts about political issues. Specifically, I monitored comments and replies on Facebook and Twitter posted by news outlets, politicians and public figures and invited people who engaged in online political hostility in response to these posts for an interview. Thus, the recruitment process were crucial to ensure ecological validity of the findings one of the key purposes of the interviews were to talk with people who actually engaged in online political hostility on mainstream platforms. The selection criteria included language that was derogatory, offensive, dehumanizing, defamatory, racist or sexist, and in some instances even incitements to violence. When potential interviewees were identified, they were contacted through the platform's private message functionalities. 345 individuals were invited for an interview, of which 7% accepted the invitation. Most were men (12% identified as women) and over 40 (mean age was 52), but they were relatively diverse in terms of political orientation and geographic region. Prior to data collection, the project was approved on March 15, 2021 by the Institutional Review Board at Aarhus University (approval number: 2021-19).

Through this design, I attempt to understand people's own rationales for their behavior — the meaning they ascribe to their behavior. A classic objection — particularly from variance-based research traditions — is that people tend to rationalize their own behavior in retrospect. While self-rationalizations definitely occur, I contend that the stories that people tell themselves are important in terms of shaping their behavior, just like identities and allegiances are important for public opinion formation (Druckman et al., 2013; Finkel et al., 2020; Huddy et al., 2015; Leeper & Slothuus, 2014). More importantly, the stories that motivate political action need not be true to shape behavior, but they are necessary to understand people's political worldviews (Hochschild, 2016). By interviewing people who engage in hostile discussions, I examine how different motivations, end goals and perceived functions of social media lead to three distinct pathways to online political hostility on mainstream social media platforms.

[1] Consult Paper A for an elaborate description of the recruitment process.

3.3 Why and when people want to regulate

Interdiction through regulation might be a viable alternative strategy for addressing online political hostility. Yet, aside from the normative discussions, attempts to regulate social media are often met by concerns that the public is divided in terms of regulating online political hostility. For instance, one poll found that 82% of Americans believe that "it would be hard to ban hate speech because people can't agree what speech is hateful" (Cato, 2017). In Paper B and Paper C, I set out to examine individual-level predictors and components of statements that shape preferences for regulation.

Paper B examines when people want to restrict online political hostility using conjoint experiments. One of the key methodological drawbacks of contemporary research on public opinion about regulating online political hostility is that it leaves individuals to make up their own inferences and stereotypes regarding the severity of the statements.[2] In an adjacent literature on support for political violence, Westwood et al. (2022) notes that when researchers "ask about general support for violence without offering context, [the researchers] leave the respondent to infer what 'violence' means," which can generate misleading inferences as "support for violence varies substantially depending on the severity of the specific violent act" (see also Druckman et al., 2022; Klar et al., 2018, for a similar argument regarding affective polarization). I argue that these very problems also apply to public opinion on regulating online political hostility. Asking people "Would you favor or oppose a law that would make it illegal to say offensive or insulting things in public about [a group]?" (Cato, 2017) leaves respondents with infinite degrees of freedom to imagine what "offensive or insulting" means. In turn, abstract questions lead to overestimating the effect of target characteristics when respondents infer higher levels of severity towards some targets. This is particularly likely if partisans are more sympathetic towards specific groups, which in turn leads to inflated partisan differences. Thus, the research design I use is less prone to overestimating both the magnitude of partisan differences and the effects of target characteristics.

In Paper B, I employ a conjoint experimental design to disentangle the effects of severity and target characteristics. Through this design, I sought to generate causal inferences by maximizing internal validity through a fully randomized experimental design. First, a key advantage

[2] See Muddiman (2021) and Skytte (2021) for notable exceptions.

of conjoint experiments is that they make it possible to disentangle the relative influence of multiple attributes—in this context, whether the target or the severity of social media posts shapes people's preferences for regulation. These two attributes are normally highly correlated in hostile posts, yet conjoint experiments make it possible to disentangle their individual effects (Bansak, Hainmueller, et al., 2021; Hainmueller et al., 2014; Sniderman, 2018; Wallander, 2009). In other words, conjoint experimental designs enable me to assess the causal effect of both the target and the severity of online political hostility in the same study. Second, a concern that often pertains to experiments is that they lack ecological validity—that is, they seem artificial and unrealistic. While using tables is the most common format for conducting conjoint experiments, I used vignettes[3] (Auerbach & Thachil, 2018; Bansak, Bechtel, et al., 2021; Bansak, Hainmueller, et al., 2021; Hainmueller et al., 2015; Huff & Kertzer, 2018) mimicking the real-world nature of social media posts. The rationale for doing so was to maximize the ecological validity of the experimental treatments (Vecchiato & Munger, 2022).

In Paper C, I turn to examining how individual differences shape opposition to regulating online political hostility. In other words, why do some people oppose regulating online political hostility? As I set out in Chapter 2, there are two major accounts. From the principled conservatism perspective, people reject regulating online political hostility because they have preferences for minimal government and support a free market of ideas (Sniderman & Carmines, 1997; Sniderman & Piazza, 1993; Sniderman et al., 1989). In other words, the opposition is shaped by principled political values stemming from conservatism. The group-based dominance perspective, in contrast, suggests that opposition to regulating online political hostility derives from desires to dominate other groups. For instance, group-based dominance is associated with racism, sexism, prejudice towards minority groups and opposition to affirmative action (Costello & Hodson, 2011; Esses et al., 2008; Ho et al., 2015; Hodson et al., 2010; Kteily et al., 2015; Rabinowitz et al., 2009; Sidanius et al., 1994; Sidanius & Pratto, 2001; Sidanius et al., 1996).

A recent study shows that social dominance orientation is associated with acceptance of hate speech (Bilewicz et al., 2017), which at face value yields support for the group-based dominance account. Yet social dominance orientation—measured through the eight-item SDO_7 scale (Ho et al., 2015)—consists of two subdimensions: dominance (SDO-D)

[3] In Paper B, I provide an example of the vignettes

and egalitarianism (SDO-E) (see also Jost & Thompson, 2000; Kugler et al., 2010; Schmitt et al., 2003). The social dominance orientation scale measures "individual differences in the preference for group-based hierarchy and inequality." The dominance subdimension (SDO-D) measures the preference for "systems of group-based dominance in which high status groups forcefully oppress lower status groups" and is associated with aggressive behaviors toward subordinate groups and endorsement of beliefs that justify oppression, including "old-fashioned racism." People who score high on the dominance subdimension would be more supportive of active and sometimes violent maintenance of status hierarchies in which some groups dominate lower status groups. The egalitarianism subdimension (SDO-E) measures a "preference for systems of group-based inequality that are maintained by an interrelated network of subtle hierarchy-enhancing ideologies and social policies" (Ho et al., 2015). People with strong anti-egalitarian values prefer "hierarchies where resources are inequitably distributed, and which can be defended by anti-egalitarian ideologies" which gives rise to preferences for minimal interference from government and libertarian free speech principles p. 1022 Ho et al., 2012 and opposition to progressive social policies and affirmative action (Jost & Thompson, 2000).

The core methodological contribution of Paper C is testing the two sub-dimensions' association with public opinion on regulating online political hostility. This in turn makes it possible to re-examine the theoretical debate between the group-based dominance perspective (Sidanius et al., 1996) and the principled conservatism perspective (Sniderman & Carmines, 1997; Sniderman et al., 1991). In other words, while re-examining the link between the subdimensions of SDO_7 in Paper C provides a methodological advance through refined measurement, the outcome yields a substantive answer to the theoretical debate.

The data for Paper B and Paper C was collected between February 2-9, 2021, when surveys with embedded conjoint experiments were fielded in two nationally representative samples in Denmark and the United States. The survey agency YouGov conducted 1518 interviews in the United States and 1535 interviews in Denmark, amounting to a full sample size of 3053 participants. Fielding the studies among nationally representative samples in Denmark and the United States makes it possible to assess the hypotheses in different contexts. The core benefit is that the cross-national samples enable me to provide inferences about lay intuitions about online political hostility across key institutional differences in legal frameworks and levels of trust and polarization in the

public. While both countries are Western liberal democracies, there are both important similarities and differences in relation to online political hostility. Key among the similarities are that civil liberties lie at the core of both countries' political cultures, and their publics are said to be some of the most supportive of free speech in the world (Skaaning & Krishnarajan, 2021; Wike, 2016).

Yet Denmark and the United States also differ in important ways, despite sharing a range of characteristics. First, Denmark stands out as a high-trust country with relatively low levels of political polarization. By contrast, political polarization has been on the rise in the United States for decades (Iyengar et al., 2019). Currently, it is heavily debated to what extent political polarization feeds into erosion of democratic norms (Broockman et al., 2022; Kingzette et al., 2021), political hostility on social media, partisan animosity and ultimately political violence (Finkel et al., 2020; Kalmoe & Mason, 2022; Westwood et al., 2022). Second, Denmark's legal framework is more restrictive in terms of free speech (Bleich, 2014) as shown in table 3.2 that shows differences in free speech limitations across Denmark and the United States. In addition to limiting speech when it incites violence or uses threats, the Danish Criminal Code's Section 266b(1) restricts speech by which "a group of people are threatened, insulted or degraded on account of their race, color, national or ethnic origin, religion, or sexual inclination." Studying public opinion on hate speech restrictions in both Denmark and the United States enables me to examine the alignment between public opinion and laws across different legal contexts. Third, Denmark and the United States are markedly different types of welfare states. The state interferes more in people's private lives in Denmark compared to the United States (Esping-Andersen, 1990). Thus, people in Denmark might be more accepting of such government interference in private lives. Fourth, the United States has a higher level of ethnic heterogeneity compared to Denmark, making characteristics such as ethnicity or religion more salient in the United States (Sniderman & Piazza, 1993). Indeed, the United States has a history that is much more characterized by conflict along ethnic and racial lines compared to Denmark, symbolized in recent years by the Black Lives Matter protests. Based on this, we should expect attitudes towards regulating online political hostility to be more polarized. In sum, Denmark and the United States share fundamental values in terms of democratic norms, but differ markedly in political, demographic and institutional ways. The United States is more politically polarized and ethnically heterogeneous, and has less government interference and fewer restrictions on civil liberties compared to Denmark. Following

the approach of previous research (Aarøe & Petersen, 2014; Jensen & Petersen, 2017), I use these differences to increase the generalizability of the public opinion findings in this dissertation.

Table 3.2: Legal criteria for hate speech-related limitations of speech in Denmark and the United States.

Severity	Denmark	United States
Incivility	-	-
Degradation Dehumanization	(✓)†	-
Incitement to violence True threats	✓	✓
Target characteristics	**Denmark**	**United States**
Race Ethnicity Religion Sexual orientation Colour	(✓)†	-
Age Disability Political beliefs Social status	-	-

† The Danish Criminal Code's section 266b(1) restricts speech with intent for wider dissemination by which "a group of people are threatened, insulted or degraded on account of their race, color, national or ethnic origin, religion, or sexual orientation".

3.4 Assessing interventions

Another strategy to address online political hostility is to mitigate its consequences. One promising mitigation strategy is to provide people with relevant competences which in turn make them feel efficacious when they face online political hostility. In Paper D, I examine the effectiveness of one such strategy, namely an intervention deployed by the Danish Health Authorities during the COVID-19 pandemic in 2021 that aimed to make people share fewer false headlines about COVID-19 in Denmark. In the paper, I set out to answer two questions: Does the intervention work, and does it make people feel efficacious or scare them? To address the first question regarding effectiveness, a three-minute video and a 15-second video from the Danish Health Authorities[4] and an accuracy nudge (Pennycook et al., 2021; Pennycook et al., 2020; Roozenbeek et al., 2021) were employed in an experimental setup and compared with a control condition. Participants were evaluated through a news-sharing task consisting of 15 real and 15 false headlines. Prior to fielding the experiments, the headlines were pretested in an independent sample. The headlines were presented one at a time in random order, and respondents were asked whether they were willing to share them.[5] A range of covariates were measured prior to the experiment, making it possible to assess potential heterogeneous effects (Rathje et al., 2022). Yet a central argument in the nudge literature is that people can be nudged to change behavior. For instance, making people attentive to accuracy makes them share more accurate headlines (Pennycook et al., 2020). Thus, if these covariates were measured just before people were exposed to the treatments, the treatments might be confounded by measures of cognitive reflection that might—unintentionally—induce people to think about accuracy. Therefore the sample was collected as a two-wave panel to avoid pre-treatment bias in assessing the interventions. In the first wave, relevant correlates of misinformation sharing were measured. This setup allowed me to assess potential heterogeneous treatment effects of the interventions. In Wave 2, the participants were only exposed to the interventions, and their willingness to share headlines across 15 real and 15 false headlines relating to COVID-19 was measured. By this procedure, the risk of confounding measurement of covariates with the experimental treatment is minimized

[4]Through collaboration with the Danish Health Authority, we have permission to use the campaign video in our studies. Yet the videos are also freely available and circulated on social media.

[5]See the experimental protocol section in Paper D for details.

by design. To address the second question, another survey experiment was fielded that included the same interventions. Yet in this experiment the effects on threat appraisal, self-efficacy and response efficacy derived from protection motivation theory were measured (Jørgensen et al., 2021; Maddux & Rogers, 1983; Rogers, 1975). In other words, the design made it possible to asses whether the interventions induced fear or feelings of competence.

Paper D sought to maximize internal validity by randomization of the treatment. In other words, the design of Paper D makes it possible to generate valid causal inferences about the effects of the interventions, both in terms of sharing of online political hostility and in terms of the feelings they induce for the individuals who are exposed to them. One caveat of the design, however, is that it does not allow for assessing the long-term effects of these interventions. In other words, the experimental setup makes it possible to assess the short-term causal effects of the interventions. Paper D also sought to maximize ecological validity by testing interventions from the Danish Health Authorities that were employed during the COVID-19 pandemic in Denmark and the extent to which they reduced the sharing of false headlines that had actually been shared on social media. In this regard, headlines were used instead of full articles (people often share headlines on social media without reading the full article) (Gabielkov et al., 2016). Overall, Paper D sought to generate valid causal inferences about the effectiveness of interventions against online political hostility in ecologically valid settings.

In summary, the four papers contributes to building a mosaic of knowledge in terms of addressing online political hostility. Paper A creates the foundation for this by interviewing people who engaged in online political hostility on mainstream social media. Specifically, this contributes to ensuring measurement and ecological validity, as I demonstrate the pathways to hostility are indeed political by recruiting people from the platforms. Paper B and Paper C rely on nationally representative samples in Denmark and the United States to assess the hypotheses in different contexts which contributes to the generalizability of the results. The key methodological contribution of Paper B relates to internal validity, as I disentangle the factors that shape support for regulating online political hostility, while Paper C provides an innovative step in terms of measuring opposition to regulating online political hostility. Finally, the main methodological contributions of Paper D are first its assessment of the effectiveness on interventions on sharing of headlines that were were widely circulated on social media — thus maximizing ecological validity — in a design that is well-equipped

to generate valid causal inferences. While these papers diverge in methodological approaches, I argue that they based on the above contribute as a whole create a mosaic of knowledge on how online political hostility can be addressed. I now turn to presenting the findings of the dissertation.

Chapter 4
Findings

This chapter summarizes the core findings of the dissertation in relation to the research question: How can online political hostility be addressed? I start out by highlighting the key findings of Paper A, in which I argue that online political hostility is a form of deliberate political participation, rather than an ”accident” that occurs on social media. Because the motivations that guide the behavior are relatively steadfast and stable, I propose addressing online political hostility by shifting attention from the producers of hostility to its audience through two strategies seeking to interdict and mitigate. Paper B and Paper C provide empirical evidence on public opinion regarding interdicting online political hostility through regulation. I show that people oppose regulating online political hostility because of principled political values, yet most people do want to restrict extreme speech, regardless of whom it targets. Paper D concerns mitigating online political hostility and presents empirical evidence that building user competences is an effective strategy to empower citizens to identify and share less online political hostility. Overall, online political hostility is a political act that can be addressed in the short-term by enforcing clear regulatory norms or bolstering competences to the audience.

4.1 Engaging in online political hostility

A good starting point for addressing online political hostility is to understand why people engage in it, and Paper A provides an empirical assessment of this. The findings are based on 25 interviews with people who engaged in online political hostility in comments sections on mainstream social media platforms in Denmark. Here, I present three of the key findings.

The first finding pertains to whether online political hostility is a political act or not. People who engage in hostility in political discussions on social media are often regarded as not having a political purpose.

Instead, they are often dismissed as ”trolls” out to ruin the internet, or else hostility is explained by features of social media that change people’s behavior for the worse (Buckels et al., 2014; Cheng et al., 2017; Rowe, 2015; Stein, 2016; Wolchover, 2012). The interviews, however, showed that people who engaged in hostility in political discussions on mainstream social media platforms did indeed hold political motivations for their behavior. In other words, the interviewees engaged in online political hostility to express their political opinions and justified their hostility on these grounds. As a consequence, people who may be seen as trolls trying to deliberately derail comments sections are actually better understood as political activists when they engage in hostility online as they perceive themselves as participating in democratic activities.

The second key finding specifically challenges the assumption that people regret their actions when they engage in hostility. Non-political accounts of online political hostility emphasize that people fail to control their emotions online, and would predict that when people cool down, they repent. However, even though the 25 interviewees were specifically recruited because they engaged in hostility in political discussions on social media, 22 of them had no regrets at all, while 3 stood by what they wrote, but said they would reconsider their framing. In other words, while most of the interviewees acknowledged or were aware that what they wrote was offensive, they justified their behavior because it expressed their political beliefs. This underlines that online political hostility is a form of deliberate political behavior.

Third, through the interviews, I identified three distinct pathways to hostility: those of the ventilator, the collider and the megaphone. The ventilator uses social media to express political frustrations and in turn to find relief. The ventilator often displays strong emotions on social media, which serves as an outlet. These emotions can be triggered by everything from disagreeable opinions to news stories or politicians. When the emotions are triggered, some people just cannot hold back as expressed by IP4: ”Then I really become. . . then I boil. I can, even now I can feel myself boiling inside as soon as I mention [female politician]”. The ventilators often described social media platforms as an arena of relief, where they could express the frustrations of daily life. Instead of sitting and yelling at the television, they are often encouraged by their partners to find relief online. One of the interviewees illustrates this point, ”Once in a while I have to let off some steam and my wife doesn’t want to listen to it . So, I directed it to Facebook. And I try to moderate myself. But I am also committed when I do something. I try to express myself as politely as I can. But there are some [times] where

it is difficult [to moderate oneself]. That is, sometimes I have to be mindful of not getting too primitive" (IP14). Overall, for the ventilator social media serves as a venue to express political frustrations and find relief. The need for relief doesn't necessarily entail a hope or belief in changing other people's minds, but rather an urge to "let off steam." Salient political issues are a key trigger in this regard, as exposure to political disagreement fuels the need to vent.

Table 4.1: Pathways to online political hostility

Type	**Motivation**	**End goal**	**Interviewees**†
Ventilator	To express frustrations	To find relief	8 (11)
Collider	To deliberate	To pursue the truth	8 (11)
Megaphone	To persuade	To gain influence	9 (11)

† Number of interviewees who were primarily categorized as each type. As some shared multiple characteristics, I included the total count in parentheses.

The collider wants to deliberate in pursuit of "the truth," but clashes with people who hold opposing viewpoints. Social media enables people to connect and share viewpoints, and colliders do exactly that. They hold genuine political opinions and want to share them in political discussions and in that sense, colliders are using social media the way it was intended. Yet, in this endeavor, they value what they perceive as facts at the expense of civility: "I am the type [...] I don't care. I stand by what is correct and what I mean. So they can call me whatever they want..." (IP16). Thus, social media works as a particle generator that causes people with markedly different perspectives to clash. As colliders are concerned with what they perceive to be the truth, they are often provoked by people who they perceive as simply trying to win an argument, despite being wrong: "I get a little tired when I come across people—and I do it fairly often—who still believe it is more important to win discussions than to clarify the realities of what we're dealing with. [...] If I wanted to have that [kind of discussion], well then I'd probably still go to the pub [as the IP previously did a lot]. That's the

level [of the debate]" (IP15). In these instances, motivations to reach accurate conclusions in political discussions prompt hostile reactions, rather than factual corrections (See Johansen et al., 2022, for a similar argument). Some colliders see themselves as suppliers of facts that can in turn enlighten other people: "You can say that it is consumer information in a way. At least, it is facts. If people are resistant to facts, then..." (IP20). Given that they share what they believe are facts, they feel that the use of hostile expressions in political discussions is justified to get those facts across. Overall, colliders engage in the ways that were envisioned in the early days of widespread internet use, when it was expected to foster more enlightened and democratic public debate. Colliders are politically interested and want to reach accurate conclusions in the political discussions they engage in. However, they sometimes end up in discussions with people who hold worldviews different from their own, and here collisions result. Thus, hostility emerges as a byproduct of political discussions, and is justified on the basis of pursuing the truth.

Social media provides a venue for sharing one's opinion, and the megaphone uses social media for exactly that purpose: trying to make themselves heard above others. Thus, social media is a tool in their endeavor of persuading others and seeking influence. Megaphones often hold strong political identities and forceful opinions on specific issues. As a consequence, they participate in social media discussions on issues they care deeply about: "I participate when people start destroying our country. I would probably call myself mostly national-patriotic in this sense. There's plenty of politicians I read posts from on which I never comment and perhaps just give them a like. But when there are a lot [of politicians] when I think that they start doing subversive activities, THEN I get involved" (IP3). The purpose of this behavior is to persuade and perhaps even mobilize and coordinate political action on social media, as expressed by IP7: "We need to get people in Denmark to wake up! Otherwise we'll simply be taken over in no time. [...] So yes, I react strongly to that, I do." The importance of sharing opinions and frustrations often makes the megaphones impervious to reactions from others, particularly because they are so committed to their worldview. Importantly, megaphones often feel empowered by social media, which is different from the ventilator, who uses social media to find relief but does not think it changes anything. The megaphone's viewpoint is distinct, because they believe that sharing their political frustrations on social media may alter political outcomes, even if it involves political hostility. Overall, for megaphones social media is a tool to gain influence by persuading other people of their opinions. These opinions are often

rooted in forceful political convictions and highly salient identities, which in turn prompt strong reactions when threatened. A reoccurring theme is to ”wake people up,” which is often pursued through strong language and discussions that sometimes lead to hostility.

In sum, Paper A shows that social media serves distinct functions based on people's political motivations, which in turn results in multiple pathways to hostility. In these instances, harsh and hostile language was perceived to be effective. Although the people interviewed provided different rationales, goals and motivations for their behavior, they were all motivated by politics and used social media to gain influence, find relief or pursue the truth. This suggests that hostility in political discussions is better understood as a political act, rather than as accidental or deliberate maladaptive behavior without a political purpose.

4.2 Addressing online political hostility

In the previous section, Paper A suggested that online political hostility is shaped by stable social and political factors which are takes time to address. In the short-term, one potential strategy is shifting the focus from the perpetrators of online political hostility to the audience. I propose that online political hostility can be interdicted by protecting the wider audience from exposure to hostility or mitigated by empowering the audience with specific competences. Here, I summarize the empirical findings of this dissertation regarding regulation and building competences among the audience to online political hostility. Paper B and Paper C constitute the basis for the findings on addressing online political hostility through regulation, while Paper D provides the empirical evidence for how the audience can be empowered.

4.2.1 Addressing online political hostility through regulation

One way to counter online political hostility is through regulation. While tech platforms and policymakers face demands to regulate content, regulating free speech in public debates is far from easy. Regulation requires prohibiting some kinds of speech. This raises concerns about limiting free speech, particularly whether these regulations can be used to silence legitimate political opposition. Can people even agree upon

what type of content that should be regulated, or do they selectively want to censor opposition for political gain?

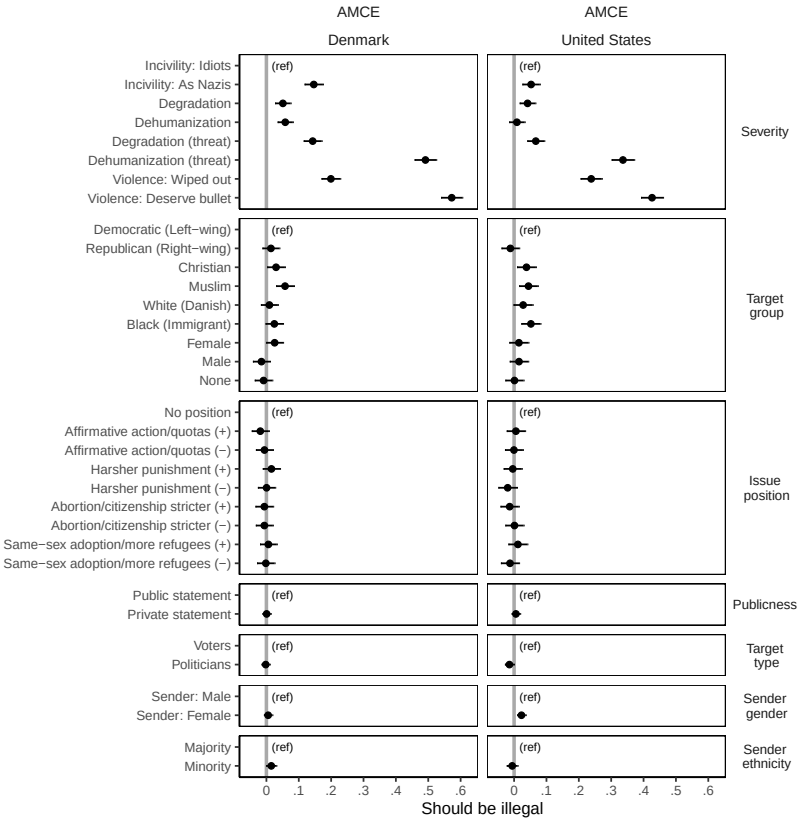
The core claim in Paper B is that people almost exclusively want to regulate online political hostility based on the severity of the content, regardless of whom it targets. In Paper B this argument is advanced through conjoint experiments in Denmark and the United States. Through these experiments, I manipulate who the hostility is targeting and the severity of its content.[1] As I highlighted in Chapter 3, Denmark and the United States diverge in their legal frameworks for regulating free speech and in terms of political polarization, making them optimal cases for drawing inferences across institutional and political contexts. Yet I find almost identical empirical patterns for Danes and Americans. The results suggest that most people are consistent in their disapproval of extreme hostility and that people do not selectively want to censor political opposition. Rather, they simply reject extreme and violent speech.

The evidence for these claims starts in Figure 4.1, in which I show the average marginal component effect (AMCE) of each attribute in terms of preferences for regulating a statement (0 = should be legal; 1 = should be illegal). The AMCE for each attribute is the marginal effect of seeing a particular attribute level in a statement relative to the baseline on the preference for restricting a statement, averaging across all other possible combinations of statement attributes (Hainmueller et al., 2014). For instance, "Democrats" is the baseline level for the "Target group" attribute in the United States. Thus, the AMCE estimate of "Republican" can be interpreted as the average causal effect on preferences for restricting a statement targeting Republicans compared to Democrats.

Figure 4.1 suggests that severity heavily shapes preferences for regulating online political hostility in both Denmark and the Untied States. The largest effects occur for statements involving threats and inciting violence. As expected, Danes and Americans are more willing to regulate extreme speech compared to incivility. In legal frameworks such as hate speech laws, target characteristics are an important criterion for limiting free speech. Based on this, we should see citizens more willing to regulate hostility directed at certain ethnic, religious and/or gender groups compared to political groups. In the panel for target groups, Americans are shown to be more supportive of regulating online political hostility targeting Christians, Muslims and Black people. However,

[1] In the conjoint experiments, seven attributes were manipulated in total. Please consult Paper B for a detailed description.

Figure 4.1: AMCEs on willingness to regulate in Denmark and the United States



Note: Figure 4.1 shows AMCEs from a regression of *severity*, *target group*, *issue position*, *publicness*, *target type*, *sender gender* and *ethnicity* on *willingness to restrict* as the dependent variable. Points are OLS estimates with 95% confidence interval bars based on clustered standard errors at the respondent level. Sample size: DK: 18655 observations (1388 respondents); US: 19130 observations (1408 respondents). See Section C in the supplementary material of Paper B for the numeric values of estimates and standard errors.

there are no statistically significant differences between preferences for regulation between these groups (e.g. people are not more willing to restrict hostility targeting Muslims compared to Christians). Similarly, in Denmark, citizens are more supportive of regulating hostility targeting Muslims and Christians, suggesting that both Danes and Americans are on average slightly more supportive of regulating online political hostility targeting certain groups. In sum, these results suggest that Americans and Danes almost exclusively rely on the severity of the content in terms of forming preferences for regulating online political hostility.[2]

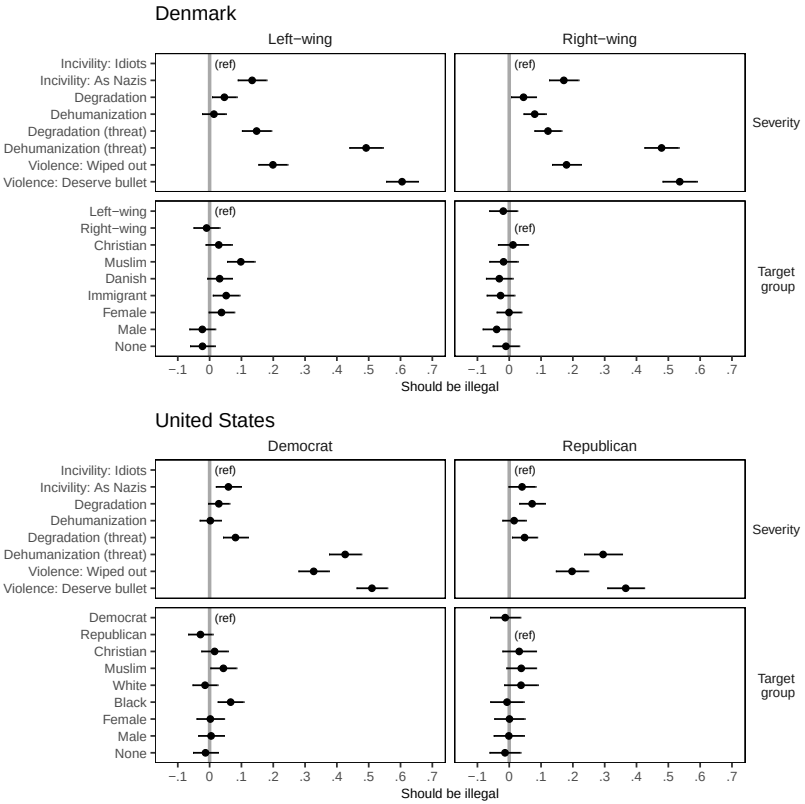
One explanation for the target having little influence on preferences for regulation is that people want to protect different groups from hostility. For instance, we might expect partisans to selectively "censor" hostility when it is targeting their own group, while turning a blind eye when it is targeting political opposition. Yet as shown in Figure 4.2, I find no evidence for this, as partisans in neither Denmark nor the United States want to regulate online political hostility selectively (cf. Ashokkumar et al., 2020; Lelkes & Westwood, 2016; Tappin & McKay, 2019).

Another proposition is that partisans simply want to protect different targets from hostility. I find supporting evidence for this in Figure 4.2, as people on the political left are more supportive of regulating online political hostility when it is aimed at certain targets. Specifically, left-wingers are more supportive of regulating hostility targeting Muslims and immigrants. While the effect for immigrants is not distinguishable from "Danes" within the ethnic category, the effect for Muslims is significantly larger than for Christians within the religious category. A similar picture emerges in the United States, as Democrats are more willing to regulate hostility targeting Black people and Muslims. Right-wingers and Republicans, in contrast, are not more or less supportive of regulating hostility based on the target. In other words, people on the political right exclusively want to regulate online political hostility based on its severity, while people on the political left are slightly more supportive of regulating online political hostility against ethnic and religious groups compared to political groups.

Overall, despite differences in hate speech legislation, I find consistent evidence that the target of the hostility matters only to a limited extent and only for some people on the political left when it is targeting certain minority groups. However, in both Denmark and

[2]Across all the other attributes, I consistently find null effects for publicness, issue positions, target type and sender ethnicity in both Denmark and the United States, with the exception of a small effect of female sender in the United States.

Figure 4.2: AMCEs on willingness to restrict across partisanship in Denmark and the United States



Note: Figure 4.2 shows AMCEs from a regression of *severity*, *target group*, *issue position*, *publicness*, *target type*, *sender gender* and *ethnicity* on *willingness to restrict* as the dependent variable. See Section C in the supplementary material of Paper B for the full regression output. Points are OLS estimates with 95% confidence interval bars based on clustered standard errors at the respondent level. Estimations are based on 14593 observations among 1062 respondents in Denmark (576 Left-wingers; 486 Right-wingers) and 15003 observations among 1076 respondents in the United States (397 Trump voters; 679 Biden voters).

the United States—and across the political spectrum—severity primarily shapes people's preferences for regulating online political hostility, regardless of whom it targets.[3] Another finding in Paper B and other studies is that people on the political left in general are more supportive of regulating online political hostility (see also Skaaning & Krishnarajan, 2021). In other words, people on the political right oppose limits on freedom of expression to a greater extent. What explains these differences?

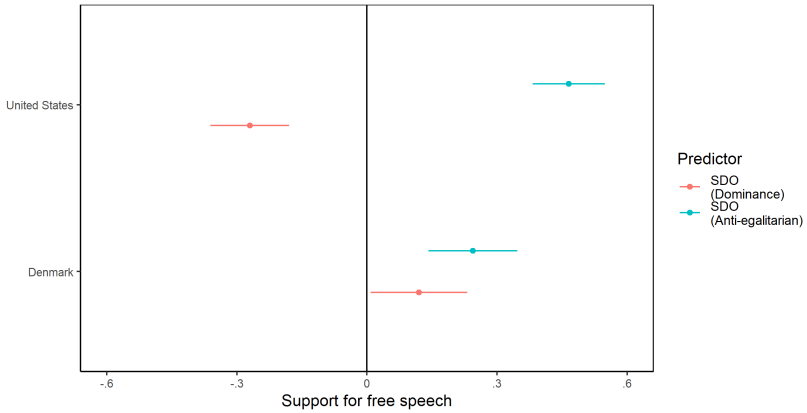
In Paper C, I test two competing perspectives on why people oppose regulating online political hostility. According to the *group-based dominance* perspective, people oppose regulating online political hostility because it is a useful tool for dominating minority and historically oppressed groups. In other words, the opposition stems from prejudice rather than ideology (Bilewicz et al., 2017; Sidanius et al., 1994; Sidanius & Pratto, 2001; Sidanius et al., 1996). The *principled conservatism* perspective, on the other hand, suggests that the opposition is based on principled political values deriving from conservatism. In other words, the opposition stems from preferences for limited government and a free market of ideas (Sniderman & Carmines, 1997; Sniderman & Piazza, 1993; Sniderman et al., 1989). Thus, the two perspectives essentially debate whether the opposition is shaped by "racism or is based instead on a principled objection to the nature" of the regulations (Feldman & Huddy, 2005).

The core contribution of Paper C is that I provide empirical support for the *principled conservatism* perspective. I arrived at this finding by examining the predictive power of the two subdimensions of social dominance orientation: dominance (SDO-D) and egalitarianism (SDO-E). In Figure 4.3 and Figure 4.4, I demonstrate that opposition to regulating free speech is shaped by anti-egalitarian values rather than group-based dominance motivations. These findings suggest that opposition to regulating online political hostility is shaped by a principled opposition to limiting free speech, rather than by selective motivations to dominate minority or historically oppressed groups. This changes the interpretation of why people oppose regulating online political hostility, emphasizing differences in political values instead of prejudice per se. [4]

[3] In the supplementary material of Paper B, I also assess the interaction between severity, target group and partisanship and find consistent evidence with this conclusion: The effects of target groups are consistently small, even for people on the political left and even when severity is high

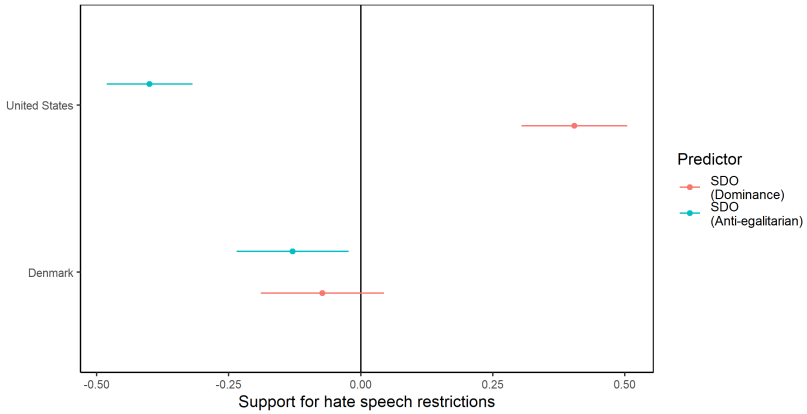
[4] Dominance motivations may still play a key role in shaping *behavior* in the form of engaging in online political hostility, which I show in Paper C, but *attitudes* towards regulating online political hostility are primarily shaped by anti-egalitarian values.

Figure 4.3: Anti-egalitarian values predict support for free speech in Denmark and the Untied States



Note: Figure 4.3 shows estimates for support of free speech for opinions that may offend or hurt other people. SDO-E and SDO-D are regressed in the same model. Estimates are based on a OLS regression with 95% confidence interval bars. Sample size: DK: 1535 respondents; US: 1518 respondents. See the supplementary material of Paper C for the numeric values of estimates and standard errors.

Figure 4.4: Anti-egalitarian values predict opposition to hate speech restrictions in Denmark and the United States



Note: Figure 4.4 shows estimates of social dominance orientation (SDO, SDO-D, SDO-E) and willingness to restrict hate speech. Estimates are based on an OLS regression with 95% confidence interval bars. Sample size: DK: 1388 respondents; US: 1408 respondents. See the supplementary material of Paper C for the numeric values of estimates and standard errors.

In summary, the results of Paper B and Paper C suggest that people oppose regulating online political hostility because of anti-egalitarian values. However, most do not selectively want to regulate online political hostility in order to further their political goals. Rather, they want to restrict extreme speech regardless of whom it targets. Meanwhile, people do indeed disagree upon regulating online political hostility, but they do so largely because of principled opposition to limiting free speech. Even across two markedly different institutional, demographic and political contexts, Danes and Americans want to restrict hate speech because of the severity of its content. These results suggest that if policymakers and tech platforms want to regulate online political hostility in line with public opinion, they should emphasize severity as the key criterion.

4.2.2 Addressing online political hostility by empowering the audience

Another way to address online political hostility is by empowering citizens by building competences to mitigate the adverse consequences of engagement with online political hostility. One of the initial hopes for social media was that it would foster a democratization of political deliberation, as anyone could engage and participate in political discussions they would not otherwise have been exposed to. However, some of the concerns that have emerged regarding social media discussions are related to the observation that people sometimes lack competences and knowledge to engage productively online. For instance, at the outset of the COVID-19 pandemic authorities raised concerns about misinformation (Zarocostas, 2020, cf. Altay et al., 2022), which constituted a threat to informed public deliberation and public health. Regulating free speech during crises is inherently problematic for numerous reasons, including undermining legitimate political opposition. In this regard, equipping citizens with competences to face the threat of misinformation during a crisis is a much more viable and less intrusive option. The question is whether it is possible to equip citizens with competences to avoid sharing online political hostility.

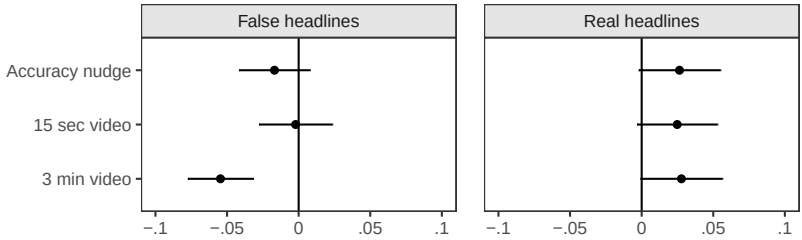
In Paper D, I examined the effectiveness of interventions against misinformation sharing during the COVID-19 pandemic. Using a set of survey experiments, I demonstrate in Figure 4.5 that a digital literacy video intervention deployed by the Danish Health Authorities reduced the sharing of false COVID-19 headlines. The video intervention provided clear, specific and actionable guidelines on why it is important avoid sharing false news as well as how to identify it and thus equipped citizens

with competences to address the risk of misinformation. Specifically, the intervention made people share fewer false headlines, but did not affect people's sharing of real headlines compared to a control condition. These results suggest that people can be empowered to share less online political hostility without affecting people's sharing of trusthworthy headlines.

Based on the evidence, I argue that people shared less online political hostility because they were empowered. However, there might be other reasons why they shared less hostile content. For instance, they may have felt scared and stopped sharing anything. In another experiment, illustrated in Figure 4.6, I demonstrate that the intervention indeed made people feel more competent in terms of not sharing COVID-19-related misinformation, without inducing fear. This suggests that public authorities can provide competences to citizens during crises.

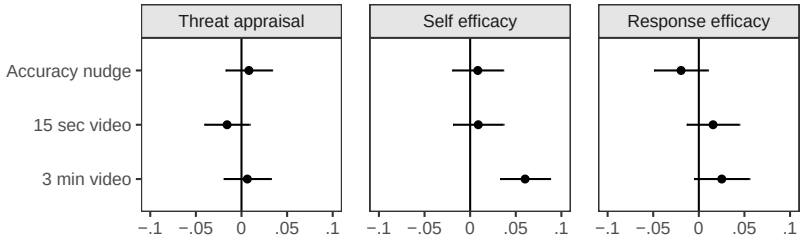
Overall, these experiments show that authorities can empower citizens engage less in online political hostility by building competences, even without affecting their sharing of trustworthy headlines or inducing fear.

Figure 4.5: Willingness to share real and false headlines



Note: Points are OLS estimates with 95% confidence interval bars based on clustered standard errors at the respondent level from two regressions: The left panel is based on a regression of the treatment conditions on the willingness to share false headlines as a dependent variable, while the right panel has the willingness to share real headlines as a dependent variable (scaled 0-1). Both regressions are based on 33,480 observations across 2,232 respondents.

Figure 4.6: Effect of interventions on threat appraisal, self efficacy and response efficacy



Note: Points are OLS estimates with 95% confidence interval bars based on clustered standard errors at the respondent level from three regressions. Each panel represents a regression of the treatment conditions on the respective pmt measure as the dependent variable. All regressions are based on samples of 2,012 respondents.

Chapter 5
Discussion

In this dissertation, I set out to answer *how online political hostility can be addressed*. In the previous section, I presented the empirical findings. In this chapter, I start by summarizing these empirical findings in a synthesised argument. I then turn to discussing the dissertation's contributions and limitations, the future directions of this research, and how it all matters.

5.1 Synthesis of findings

In this dissertation, I have argued that instances of hostility in political discussions on social media are political acts. Although I showed that there were numerous pathways to hostility, people were not hostile in political discussions online by accident—they used social media to convey political frustrations, facts and opinions to a wide audience. In this way, the affordances of social media function as a vehicle for numerous political goals, such as persuading others, ventilating or discussing politics. Because these acts are motivated by political beliefs and frustrations from the real world, the motivations has to be addressed in the offline world, as subtle nudges do not address to root of the problem. For instance, extreme mistrust or feelings of marginalization are largely stable and require structural changes, which are demanding and take time. Yet policymakers and social media platforms face demands to do something, because many people experience hostility in online environments — particularly in political discussions on mainstream platforms. In other words, there is a demand for solutions while waiting for policies that address the root causes, because social media is one of the primary venues for political discussions. What can policymakers and social media platforms do — in the short-term — to address online political hostility?

I outlined two ways to address online political hostility using regulation and empowerment, respectively. Essentially, these measures

shift the perspective from the perpetrators of online political hostility to its "innocent" audience and reach quite optimistic conclusions. First, they suggest that the public is quite consistent in its preferences for regulating online political hostility, and bases these preferences on political values. In other words, people don't just want to censor content on social media for political gain. Second, the results suggest that authorities can empower citizens when they encounter online political hostility by providing them with specific competences. In short, the public is more aligned when it comes to free speech regulation and can be empowered more than previously thought. In the following, I discuss the contributions and implications of the empirical findings and outline avenues for future research.

5.2 Contributions, limitations and future directions

5.2.1 Politics matters

The key contribution of Paper A is to show that politics matters. As I outlined in Chapter 2, extensive literatures have suggested that hostility in political discussions on social media reflects bad personalities or the affordances of social media which in turn inform the way online political can be addressed. These highlight how trolls ruin political discussions on social media either deliberately (Buckels et al., 2014; Stein, 2016) or because of the design of social media (Cheng et al., 2017; Wolchover, 2012). The empirical findings in Paper A contrast this narrative. I do not dispute that affordances of social media or certain personality traits may fuel online political hostility—quite the opposite. However, I do contend that the 25 interviewees I talked to, all of whom had previously engaged in hostility in political discussions on mainstream social media platforms, were fueled by political motivations. In other words, I concur with the argument that people who have a hard time controlling their emotions or have sadistic personality traits engage in more hostility; however, I argue that the hostility on social media emerges in the interplay with political motivations, which in turn leads to distinct pathways for engaging in online political hostility. While it is up to future research to assess the generalizability of these results, one direct implication of this finding is that a lot of the hostility on social media reflects genuine political frustrations that found an outlet on social media.

The three pathways I identified have implications for how online political hostility can be addressed. First, the ventilators use social media as a speaker's corner in which they can vent their frustrations. These frustrations are fueled by lack of trust in politicians or institutions, which seems to generate political cynicism. The past several decades have been marked by increasing political polarization and inequality and decreased trust (see e.g. Iyengar et al., 2019). More recently, the psychological burden of the COVID-19 pandemic fueled anti-democratic sentiments (Bartusevičius et al., 2021) and stress (Kowal et al., 2020; Lieberoth et al., 2021). Addressing these developments requires a change of policies and behaviors from governments and authorities. For instance, governments can foster trust through transparent communication, even during crises (Petersen et al., 2021). Second, to address hostility from colliders would require a change of procedures for having political discussions on social media. One potential is fostering norms that are conducive to public deliberation. A vast majority of the interviewees indicated that they were attentive to both formal and informal norms in discussions on social media and deployed strategies to adapt or circumvent these norms. Further research should examine the viability of altering behavior — including online political hostility — through norms and content moderation, for better or worse. Finally, decreasing polarization—perceived or real—might be the most viable way to address the hostility of megaphones. Previous research has shown that exposure to opposing views can increase political polarization (Bail et al., 2018). Recently, however, a range of interventions showed promising results in terms of reducing partisan animosity through highlighting common identities and facilitating trust or contact (Hartman et al., 2022).

Future research could address at least two unanswered questions. The first question regards the prevalence of these pathways to online political hostility. While Paper A outlines distinct pathways, future research could assess the relative prevalence of the pathways by measuring the political motivations of people who engage in online political hostility. Another question relates to the types of online political hostility that the pathways lead to. One plausible assumption is that those I refer to as ventilators — people who engage in online political hostility to find relief from their political frustrations — are triggered by situational features when they are exposed to certain political content. Thus their hostility might be more sporadic. On the other hand, megaphones' attempts to persuade others may foster repetitive patterns in which they frequent the same environments, which

escalates discussions. Yet these hypotheses are empirical questions that require testing.

Overall, I argue that hostility in political discussions online reflects genuine political motivations and frustrations. A direct consequence thereof is that the measures required to address it are long-term and often at a structural level. The findings in this dissertation suggest that there is no quick fix because hostility on social media is a reflection of people's motivations, frustrations and grievances in their daily lives, which in turn may be amplified and become visible on social media. In this context, it is important to highlight that some of the political frustrations most likely have always been there even prior to the advent of social media. While political frustrations 20 years ago may have been widely shared within individuals social network in their offline life, they are now visible to a wider audience for better or for worse. In liberal democracies, sharing political frustrations in public — within the realm of the law — is an inherent feature of public debates, yet when the behavior undermine democratic norms it often warrants action. Thus, as the source of online political hostility is shaped by stable social and political factors that are hard to address, countermeasures in the short term should focus on either protecting the wider audience from exposure to hostility or empowering them with specific competences. In the next two sections, I turn to two ways of addressing online political hostility that either protect the wider audience through regulation or provide them with relevant competences when they face it.

5.2.2 Room for agreement

Regulating content on social media is not an easy task. People believe it is hard to ban hate speech because they think people cannot agree upon what hate is (e.g. Cato, 2017; Dunn, 2019). Most Americans believe social media censors political viewpoints (Vogels et al., 2020) and people on either side of the political spectrum are divided on whether offensive content online is taken seriously enough (a Vogels, 2020). In this dissertation, I have argued that people essentially agree that online political hostility should be regulated on the basis of severity, while opposition to regulating online political hostility in general is shaped by anti-egalitarian political values. What do these results mean for our previous knowledge, and what implications do they have for regulating online political hostility on social media platforms?

The results of this dissertation stand in contrast to the narrative that people are deeply divided on what types of content should be restricted

because people are intolerant of groups and ideas they dislike (Gibson, 2011; Marcus et al., 1995; Stouffer, 1955). Some warn that people prioritize political goals over democratic norms (Frederiksen, 2022; Graham & Svolik, 2020; Simonovits et al., 2022; Svolik, 2018) and that if people had the chance, they would be willing to censor their political opponents (Amira et al., 2021; Ashokkumar et al., 2020; Lelkes & Westwood, 2016). In contrast, I argue in Paper B that people—including those with distinct political allegiances—consistently want to regulate online political hostility based on the severity of its content. Yet, as shown in Paper C, people do disagree upon the threshold for regulating online political hostility — but they do so based on principled opposition to restricting free speech.

Another key contribution of the dissertation is the use of conjoint experiments to disentangle public opinion on regulating online political hostility. Previous research tends to conflate two of the key dimensions of online political hostility — its severity and its target — which in turn makes it impossible to differentiate their effects (e.g. Cato, 2017). Drawing inferences based on abstract question wordings may lead researchers astray in two important ways. First, they may — mistakenly — infer that people are more willing to regulate hostility directed at one group, when people in reality merely infer more severity based on ambiguous question wordings that leave it up to the respondent to ascribe the level of severity (e.g. by asking "Would you favor or oppose a law that would make it illegal to say offensive or insulting things in public about [a group]?"). Second, the ambiguous question wordings may overestimate both the level of support for regulation and the partisan difference, because they do not provide proper context and thus give respondents infinite degrees of freedom to interpret the questions (see e.g Druckman et al., 2022; Klar et al., 2018; Westwood et al., 2022). In sum, the use of conjoint experiments contributes to our understanding of public opinion on regulating online political hostility by disentangling the factors that shape support for regulation and generating more precise measures through concrete questions. These are not just trivial methodological advances, because they indeed do provide a more optimistic account of regulating online political hostility by suggesting room for agreement. Based on this, the real challenge is finding common ground on the severity threshold for regulating online political hostility on social media. Thus, the straightforward implication in relation to regulating online political hostility is that online community standards should emphasize when content violates guidelines based on

its severity rather than target characteristics, because severity shapes public opinion on regulating speech.

Future research should examine the consequences of different types of content moderation. In its broadest sense, content moderation can happen in multiple ways including through legislation from authorities, community standards on social media platforms, or crowd moderation where users flag hostile content. However, these efforts should be informed by the argument people essentially agree online political hostility should be regulated on the basis of its severity, although people have different thresholds for limiting free speech. A direct implication of this finding is that emphasizing severity in content moderation makes guidelines clearer, more transparent and in line with public opinion. People follow rules when they believe that authorities are legitimate, which is derived from a notion that people are treated fairly and equally (Tyler, 2006b). In turn, when people internalize social norms and values, they regulate their own behavior in line with the rules (Tyler, 2006a). Indeed, previous research highlights how content moderation systems on mainstream platforms remove online political hostility at scale, but do little to educate citizens on where they went wrong through transparent communication (Myers West, 2018). While some people might disagree that their content should be moderated, some evidence suggests that transparent public communication disclosing negative information facilitates trust (Petersen et al., 2021). In sum, if social media platforms are to foster environments with sustainable rule enforcement, they should emphasize clear and transparent guidelines that are in line with public opinion. Future research should examine the effects of content moderation that provides clear and transparent communication about why content was removed and outline guidelines for acceptable behavior focusing on content severity.

An alternative to formal content moderation is informal norm enforcement by witnesses to online political hostility—often referred to as bystanders. In social contexts, people are attentive to norms and rules. This is particularly true when they are enforced by people who are like them and hold high status (Hogg, 2016). A shared sense of social identity motivates prosocial behavior, and harnessing people's social identities can thus shape norms for the good (Van Bavel & Packer, 2021). Recent evidence shows that when people are "corrected" by bystanders with high status who share their identity, they engage in less online political hostility (Hangartner et al., 2021; Munger, 2016; Siegel & Badaan, 2020). Other research shows that the presence of a high number of bystanders sharing a common identity is effective in generating norms

and thus reducing hostile behavior (Paluck & Green, 2009; Paluck et al., 2021). Thus, online political hostility can be reduced by the informal engagement of bystanders. These findings are promising because while people often encounter online political hostility on mainstream platforms, most people do not act (Andresen et al., 2022). Activating bystanders thus provides a less intrusive means to reduce online political hostility. Similar to formalized content moderation, future research should scrutinize the consequences of bystander reactions. Are bystander reactions effective in terms of altering the behavior of the offender, does it encourage other bystanders to react, how does it affect the victims, and do prosocial bystander reactions in turn foster better environments for political discussions on mainstream social media platforms?

While norms can be enforced through formalized content moderation or informal bystander interventions, a general point that applies to both is that future research should assess the short- and long-term effects of these in terms of addressing online political hostility on mainstream social media platforms. These effects can be assessed either by field experiments (Mosleh et al., 2021) or by connecting experiments embedded in panel surveys to social media data and in turn assessing the persistence of the effects over time (Carey et al., 2022; Maertens et al., 2021). In this regard, reducing online political hostility is one of many outcomes relevant for public deliberation, which also includes reducing group-based animosity and polarization and fostering well-being.

5.2.3 Power to the people

While it is clear that people experience hostility online, much less is known about how to mitigate it. The literature on interventions to improve social media is dramatically expanding. The types of interventions are diverse and include fact-checking (Walter et al., 2020), corrections (Bode & Vraga, 2018) and nudges (Pennycook et al., 2021), as well as inoculation (Van Der Linden, 2022), identity (Pretus et al., 2022), digital literacy or bystander (Hangartner et al., 2021; Munger, 2016; Siegel & Badaan, 2020) interventions (see also Van Bavel, Harris, et al., 2021, for an overview). In this dissertation, however, I have argued that online political hostility is motivated by relatively stable factors such as personality traits, trust, political convictions and deeply held frustrations. Thus, there is no quick fix. Yet the findings of Paper D suggest that during times of crisis—including at the onset of a pandemic—it is possible to equip citizens with tools that both make citizens feel competent and in turn actually make them spread less

misinformation. In other words, given that people's motivations are quite stable, we need to provide specific competences through transparent communication that provides concrete, detailed and actionable advice.

Based on this, I suggest that further research examine how interventions can be used during times of political turmoil and crisis. In line with the argument in this dissertation that the focus of interventions could be changed from targeting the perpetrator to targeting the audience, I suggest that providing people with relevant competences might be a viable way to address online political hostility. Such interventions could examine whether equipping people with concrete tools leads them to share less hostility, even during heated political debates. Similarly, we might expect that equipping individuals with tools to intervene when they see hostility in online environments might reduce hostility and foster democratic norms in online environments. Surely, interventions of this sort will not provide a full solution to the downsides of political discussions in online environments, yet they might mitigate some of the adverse consequences of online political hostility.

5.3 Concluding remarks

The starting point of this dissertation was the shattered hopes of social media regarding democratic optimism. The prevalence of hostility on social media shape demands for addressing social media. But what can policymakers and social media platforms do? One of the core claims in this dissertation is that online political hostility is a political problem, rather than a social media problem. Social media provides a venue in which people can share their frustrations, discuss with other people and try to persuade them, and online political hostility emerge figuratively speaking as a form of collateral damage of this process.

The dissertation contributes by highlighting the importance of assessing the assumptions that motivate online political hostility in order to address it effectively. The ability to express one's political frustrations — within the realm of the law — is an inherent feature of democracy. Thus, one of the core tasks for policymakers and social media platforms is striking a balance where they on the one hand provide room for people to express and discuss their political frustrations and on the other hand securing that some users on social media platforms does not undermine democratic norms in political discussions. Based on the evidence of this dissertation, the long-term strategy for addressing online political hostility is to address the political frustrations of people's daily lives. Yet

hostility is already prevalent on social media, and one part of the solution might require mitigating or interdicting online political hostility in the short-term, because public debates to find solutions take place online. To this end, I proposed shifting focus from the perpetrators to how online political hostility can be mitigated or interdicted through building competences among the audience or regulated. In this regard, I believe the results of this dissertation is quite optimistic — although they neither address the root cause — as they suggest that people are less divided on regulating online political hostility than previously thought and can equipped with competences to mitigate the adverse consequences of engagement with online political hostility.

Summary

Social media inherently holds democratic potential, as it enables anyone to access information, connect with others and participate in political deliberation. Social media is often referred to as an open public square that gives voice to democratic forces—yet when people engage in political discussions on social media, they often encounter behaviors that are corrosive to democratic norms. Public deliberation relies on norms of free and open debate in which informed decisions are based on accurate information. As a consequence, policymakers and social media companies face pressure to address behaviors on social media that are actively hostile towards those norms. Such behaviors include the sharing of misinformation that undermines decision-making based on rational and informed public deliberation, or the use of hate speech that threatens free and equal participation in discussions. In this dissertation, I examine how these behaviors — which I refer to as online political hostility — can be addressed?

I start by assessing a key assumption in some of the most dominant explanations of and interventions against online political hostility: that people who are hostile in political discussions on social media are not motivated by politics. Rather, people are thought to be triggered by certain features of social media or of their personalities. In other words, some of the most widespread strategies to address online political hostility assume that it is largely apolitical and thus can be "corrected" if only people are nudged in the right direction. I argue, in contrast, that online political hostility is deliberate political acts grounded in political values and motivations. As political motivations are relatively stable, they are hard to address in the short term. Therefore, I propose shifting the focus of addressing online political hostility from the perpetrators to the audience. I assess two strategies that address the adverse consequences of online political hostility: interdiction through regulation and mitigation by empowering the audience. First, I show that while principled political values shape opposition to regulating online political hostility, people do agree that severity is the key criterion for regulation. Second, empowering citizens by equipping them with concrete tools and

advice makes them feel more efficacious in terms of both facing and engaging with online political hostility.

This dissertation show that some people who partake in political discussions online use social media as an arena where they can vent, discuss or try to persuade others. Sometimes these political motivations lead to online political hostility, motivated by frustrations and grievances from people's offline lives. Because these acts are motivated by political beliefs and frustrations from the real world, they are hard to change and require long-term policy change to remedy. Yet in the meantime I provide evidence that there is more room for agreement than previously thought in terms of regulating social media, and that people can be empowered to face online political hostility.

Dansk Resumé

Sociale medier giver adgang til information samt mulighed for at komme i kontakt med andre borgere og deltage i politiske diskussioner. I de sociale mediers barndom blev de anset som et offentligt rum med plads til en demokratisk samtale, hvor holdninger og synspunkter frit kunne udveksles. Men når folk deltager i politiske diskussioner på de sociale medier, bliver de ofte mødt med had, der underminerer demokratiske normer. En offentlig demokratisk samtale afhænger af, at der er fri og lige adgang til debatter, hvor beslutninger bliver taget på baggrund af et oplyst grundlag og faktuel information. Af samme årsag er der et stigende pres på lovgivere og de sociale medier for at gøre noget ved den adfærd, der underminerer den demokratiske samtale på de sociale medier. Denne adfærd indebærer deling af falske nyheder, der truer, hvorvidt beslutninger bliver taget på et rationelt og informeret grundlag gennem offentlig debat, og hadefulde ytringer, som truer den frie og lige adgang til diskussioner på sociale medier. I denne afhandling undersøger jeg, hvordan disse typer af adfærd – som jeg bredt referer til som online politisk had – kan imødegås.

Jeg starter med at undersøge en af de udbredte antagelser, der ligger bag nogle af de mest fremtrædende forklaringer på og interventioner mod had på nettet: At folk, der er hadefulde i politiske diskussioner på sociale medier, ikke er motiveret af politik. Ifølge disse apolitiske forklaringer bliver hadet udløst af den måde, sociale medier er designet, eller folks personlighed. Med andre ord så antager nogle af de meste fremtrædende strategier til at håndtere online had, at hadet er apolitisk og dermed kan korrigeres, hvis folk blot bliver "nudget" i den rigtige retning, eller vi ændrer den måde platformene fungerer på. I modsætning hertil viser jeg, at online had er en distinkt form for politisk adfærd og aktivisme for folk, der deltager i diskussioner på sociale medier. Fordi motivationerne er politiske, kræver de politiske eller sociale forandringer i det virkelige liv, og er svære at gøre noget ved på kort sigt. En anden mulighed er helt grundlæggende at skifte fokus fra dem, der udtrykker hadet, til det uskyldige publikum. I denne afhandling undersøger jeg to strategier, som sigter mod at håndtere de negative konsekvenser af online

politisk had gennem regulering eller opøvelse af kompetencer hos dem, der er eksponeret til hadet. Først viser jeg, at mens principielle politiske værdier former modstand mod regulering af online had, så er folk enige om, at online had skal reguleres på baggrund af dets grovhed, uagtet hvem det er rettet mod. For det andet viser jeg, at man kan mindske nogle af de negative konsekvenser af online had ved at give konkrete redskaber og råd til publikum, hvilket både øger deres tiltro til egne evner, men også får dem til at dele mindre had.

Overodnet viser denne afhandling, at folk, der er hadefulde i politiske diskussioner på social medier, bruger kommentarspor som en arena, hvor de kan ventilere, diskutere eller forsøge at overbevise andre om deres politiske synspunkter. Nogle gange fører denne adfærd til online had, som helt grundlæggende er motiveret af politiske frustrationer fra folks dagligdag. Med andre ord er motivationerne rodfæstede og kræver langsigtede ændringer for at blive afhjulpet. Mens vi venter på politiske forandringer, som kan adressere disse grundlæggende utilfredsheder, viser jeg, at folk faktisk er enige om, at online had skal reguleres på baggrund af grovheden, mens publikum til hadet kan tilegne sig kompetencer, der kan mindske de negative konsekvenser af online had.

Bibliography

Aarøe, L., & Petersen, M. B. (2014). Crowding out culture: Scandinavians and americans agree on social welfare in the face of deservingness cues. *The Journal of Politics*, *76*(3), 684–697. https://doi.org/10.1017/s002238161400019x

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Altay, S., Kleis Nielsen, R., & Fletcher, R. (2022). Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *Journal of Quantitative Description: Digital Media*, *2*. https://doi.org/10.51685/jqd.2022.020

Amira, K., Wright, J. C., & Goya-Tocchetto, D. (2021). In-Group Love Versus Out-Group Hate: Which Is More Important to Partisans and When? *Political Behavior*, *43*(2), 473–494. https://doi.org/10.1007/s11109-019-09557-6

Andresen, M. J., Karg, S. T. S., Rasmussen, S. H. R., Pradella, L., Rasmussen, J., Lindekilde, L., & Petersen, M. B. (2022). *Danskernes oplevelse af had på sociale medier*.

Ashokkumar, A., Talaifar, S., Fraser, W. T., Landabur, R., Buhrmester, M., Gómez, Á., Paredes, B., & Swann, W. B. (2020). Censoring political opposition online: Who does it and why. *Journal of Experimental Social Psychology*, *91*, 104031. https://doi.org/10.1016/j.jesp.2020.104031

Auerbach, A. M., & Thachil, T. (2018). How clients select brokers: Competition and choice in india's slums. *American Political Science Review*, *112*(04), 775–791. https://doi.org/10.1017/s000305541800028x

a Vogels, E. (2020). *Partisans in the U.S. increasingly divided on whether offensive content online is taken seriously enough*. Pew Research Center. Retrieved December 13, 2022, from https://www.pewresearch.org/fact-tank/2020/10/08/partisans-in-the-u-s-increasingly-divided-on-whether-offensive-content-online-is-taken-seriously-enough/

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization (2018/08/30). *Proc Natl Acad Sci U S A*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bansak, K., Bechtel, M. M., & Margalit, Y. (2021). Why Austerity? The Mass Politics of a Contested Policy. *American Political Science Review*, *115*(2), 486–505. https://doi.org/10.1017/s0003055420001136

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2021, April 1). Conjoint Survey Experiments. In J. Druckman & D. P. Green (Eds.), *Advances in Experimental Political Science* (1st ed., pp. 19–41). Cambridge University Press. https://doi.org/10.1017/9781108777919.004

Bartels, L. M. (2020). Ethnic antagonism erodes Republicans' commitment to democracy. *Proceedings of the National Academy of Sciences*, *117*(37), 22752–22759. https://doi.org/10.1073/pnas.2007747117

Bartusevičius, H., Bor, A., Jørgensen, F., & Petersen, M. B. (2021). The Psychological Burden of the COVID-19 Pandemic Is Associated With Antisystemic Attitudes and Political Violence. *Psychological Science*, 09567976211031847. https://doi.org/10.1177/09567976211031847

Bilewicz, M., Soral, W., Marchlewska, M., & Winiewski, M. (2017). When authoritarians confront prejudice. differential effects of SDO and RWA on support for hate-speech prohibition. *Political Psychology*, *38*(1), 87–99. https://doi.org/10.1111/pops.12313

Bleich, E. (2014). Freedom of expression versus racist hate speech: Explaining differences between high court regulations in the USA and europe. *Journal of Ethnic and Migration Studies*, *40*(2), 283–300. https://doi.org/10.1080/1369183X.2013.851476

Bliuc, A.-M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, *87*, 75–86. https://doi.org/10.1016/j.chb.2018.05.026

Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication*, *33*(9), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

Bor, A., & Petersen, M. B. (2021). The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis. *American Political Science Review*, 1–18. https://doi.org/10.1017/S0003055421000885

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Broockman, D. E., Kalla, J. L., & Westwood, S. J. (2022). Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not. *American Journal of Political Science*. https://doi.org/10.1111/ajps.12719

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97–102. https://doi.org/10.1016/j.paid.2014.01.016

Carey, J. M., Guess, A. M., Loewen, P. J., Merkley, E., Nyhan, B., Phillips, J. B., & Reifler, J. (2022). The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. *Nature Human Behaviour*, *6*(2), 236–243. https://doi.org/10.1038/s41562-021-01278-3

Carey, J. M., Helmke, G., Nyhan, B., Sanders, M., & Stokes, S. (2019). Searching for Bright Lines in the Trump Presidency. *Perspectives on Politics*, *17*(3), 699–718. https://doi.org/10.1017/s153759271900001x

Cassese, E. C. (2021). Partisan Dehumanization in American Politics. *Political Behavior*, *43*(1), 29–50. https://doi.org/10.1007/s11109-019-09545-w

Cato. (2017). *The State of Free Speech and Tolerance in America*. Cato Institute. Retrieved April 12, 2021, from https://www.cato.org/survey-reports/state-free-speech-tolerance-america

Chambers, S. (2003). Deliberative Democratic Theory. *Annual Review of Political Science*, *6*(1), 307–326. https://doi.org/10.1146/annurev.polisci.6.121901.085538

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll. https://doi.org/10.1145/2998181.2998213

Costello, K., & Hodson, G. (2011). Social dominance-based threat reactions to immigrants in need of assistance. *European Journal of Social Psychology*, *41*(2), 220–231. https://doi.org/10.1002/ejsp.769

Crosier, B. S., Webster, G. D., & Dillon, H. M. (2012). Wired to connect: Evolutionary psychology and social networks. *Review of General Psychology*, *16*(2), 230–239. https://doi.org/10.1037/a0027919

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2022). (Mis)estimating Affective Polarization. *The Journal of Politics*, 000–000. https://doi.org/10.1086/715603

Druckman, J. N., Peterson, E., & Slothuus, R. (2013). How elite partisan polarization affects public opinion formation. *American Political*

Science Review, *107*(1), 57–79. https://doi.org/10.1017/s0003055412000500

Duggan, M. (2017). *Online harassment 2017*. Pew Research Center.

Dunn, A. (2019, June 19). *6. The challenge of knowing what's offensive*. Pew Research Center - U.S. Politics & Policy. Retrieved January 24, 2023, from https://www.pewresearch.org/politics/2019/06/19/the-challenge-of-knowing-whats-offensive/

Easton, D. (1965). A Framework for Political Analysis. *Englewood Cliffs, HJ: Prentice-Hall*.

Eberwein, T. (2019). "trolls" or "warriors of faith"? *Journal of Information, Communication and Ethics in Society*, *18*(1), 131–143. https://doi.org/10.1108/jices-08-2019-0090

EIU. (2020). *Democracy Index 2019*. Retrieved November 16, 2022, from https://pages.eiu.com/rs/753-RIQ-438/images/Democracy%20Index%202019.pdf?mkt_tok=NzUzLVJJUS00MzgAAAGIIP5DNB-e-zWzV1eqJtOlv-0h0NgkVkrFhYI8dle4aaMGsu3gu70X_fvqRoIbSkX8RbWn8MF9K8DEb4XfBaWSyNHWWkNaqdIMc_CeA4zkKQpq

Erjavec, K., & Kovačič, M. P. (2012). "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920. https://doi.org/10.1080/15205436.2011.619679

Esping-Andersen, G. (1990). The Three Political Economies of the Welfare State. *International Journal of Sociology*, *20*(3), 92–123. Retrieved November 16, 2022, from https://www.jstor.org.ez.statsbiblioteket.dk:2048/stable/20630041

Esses, V. M., Veenvliet, S., Hodson, G., & Mihic, L. (2008). Justice, Morality, and the Dehumanization of Refugees. *Social Justice Research*, *21*(1), 4–25. https://doi.org/10.1007/s11211-007-0058-4

Fangen, K., & Holter, C. R. (2019). The battle for truth: How online newspaper commenters defend their censored expressions. *Poetics*, 101423. https://doi.org/10.1016/j.poetic.2019.101423

Faulkner, N., & Bliuc, A.-M. (2016). 'it's okay to be racist': Moral disengagement in online discussions of racist incidents in australia. *Ethnic and Racial Studies*, *39*(14), 2545–2563. https://doi.org/10.1080/01419870.2016.1171370

Federico, C. M., & Sidanius, J. (2002). Racism, ideology, and affirmative action revisited: The antecedents and consequences of "principled objections" to affirmative action. *Journal of Personality and Social Psychology*, *82*(4), 488–502.

Feldman, S., & Huddy, L. (2005). Racial Resentment and White Opposition to Race-Conscious Programs: Principles or Prejudice?

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., Mason, L., McGrath, M. C., Nyhan, B., Rand, D. G., Skitka, L. J., Tucker, J. A., Van Bavel, J. J., Wang, C. S., & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*(6516), 533–536. https://doi.org/10.1126/science.abe1715

Frederiksen, K. V. S. (2022). Does Competence Make Citizens Tolerate Undemocratic Behavior. *Forthcoming in American Political Science Review*.

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social Clicks: What and Who Gets Read on Twitter? Retrieved April 9, 2022, from https://hal.inria.fr/hal-01281190

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2016). *Countering online hate speech*. UNESCO.

Gibson, J. L. (2011). Political Intolerance in the Context of Democratic Theory. *The Oxford Handbook of Political Science*. https://doi.org/10.1093/oxfordhb/9780199604456.013.0021

Graham, M. H., & Svolik, M. W. (2020). Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States. *American Political Science Review*, *114*(2), 392–409. https://doi.org/10.1017/s0003055420000052

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Grossman, L. (2010). Person of the Year 2010: Mark Zuckerberg, [magazine]. *Time*. Retrieved January 20, 2023, from https://content.time.com/time/specials/packages/article/0,28804,2036683_2037183_2037185-7,00.html#

Guess, A. M., Barberá, P., Munzert, S., & Yang, J. (2021). The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, *118*(14), e2013464118. https://doi.org/10.1073/pnas.2013464118

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 201920498. https://doi.org/10.1073/pnas.1920498117

Guess, A. M., & Lyons, B. A. (2020, August 31). Misinformation, Disinformation, and Online Propaganda. In N. Persily & J. A. Tucker (Eds.), *Social Media and Democracy* (1st ed., pp. 10–33). Cambridge University Press. https://doi.org/10.1017/9781108890960.003

Haidt, J., & Rose-Stockwell, T. (2019). *Social Media Is Warping Democracy - The Atlantic*. Retrieved January 9, 2023, from https://senatus.nethttps://senatus.net/shareditems/14827/url/

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior (2015/02/04). *Proc Natl Acad Sci U S A*, *112*(8), 2395–400. https://doi.org/10.1073/pnas.1416587112

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, *22*(1), 1–30. https://doi.org/10.1093/pan/mpt024

Hangartner, D., Dinas, E., Marbach, M., Matakos, K., & Xefteris, D. (2019). Does Exposure to the Refugee Crisis Make Natives More Hostile? *American Political Science Review*, *113*(2), 442–455. https://doi.org/10.1017/S0003055418000813

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), e2116310118. https://doi.org/10.1073/pnas.2116310118

Hartman, R., Blakey, W., Womick, J., Bail, C. A., Finkel, E., Schroeder, J., Sheeran, P., Van Bavel, J. J., Willer, R., & Gray, K. (2022, February 17). *Interventions to Reduce Partisan Animosity* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/ha2tf

Hersh, E. D. (2017). Political Hobbyism: A Theory of Mass Behavior.

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science*, *12*(6), 973–986. https://doi.org/10.1177/1745691617702496

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R., & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO7 scale. *Journal of Personality and Social Psychology*, *109*(6), 1003–1028. https://doi.org/10.1037/pspi0000033

Ho, A. K., Sidanius, J., Pratto, F., Levin, S., Thomsen, L., Kteily, N., & Sheehy-Skeffington, J. (2012). Social Dominance Orientation: Revisiting the Structure and Function of a Variable Predicting Social and Political Attitudes. *Personality and Social Psychology Bulletin*, *38*(5), 583–606. https://doi.org/10.1177/0146167211432765

Hochschild, A. R. (2016). *Strangers in their own land: Anger and mourning on the american right*. The New Press.

Hodson, G., Rush, J., & MacInnis, C. C. (2010). A joke is just a joke (except when it isn't): Cavalier humor beliefs facilitate the expression of group dominance motives. *Journal of Personality and Social Psychology*, *99*(4), 660–682. https://doi.org/10.1037/a0019627

Hogg, M. A. (2016). Social identity theory. In S. McKeown, R. Haji, & N. Ferguson (Eds.), *Peace Psychology Book Series* (pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-29869-6_1

Horowitz, D. L. (2001). *The deadly ethnic riot*. Univ of California Press.

Huddy, L., Mason, L., & Aarøe, L. (2015). Expressive partisanship: Campaign involvement, political emotion, and partisan identity. *American Political Science Review*, *109*(1), 1–17. https://doi.org/10.1017/s0003055414000604

Huff, C., & Kertzer, J. D. (2018). How the public defines terrorism. *American Journal of Political Science*, *62*(1), 55–71. https://doi.org/10.1111/ajps.12329

International Covenant on Civil and Political Rights, 19 December 1966, 999 UNTS 171, Can TS 1976 No 47 (Entered into Force 23 March 1976) [ICCPR]. (1966). https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-4&chapter=4&clang=_en

Ihlebæk, K. A., & Holter, C. R. (2021). Hostile emotions: An exploratory study of far-right online commenters and their emotional connection to traditional and alternative news media. *Journalism*, *22*(5), 1207–1222. https://doi.org/10.1177/1464884920985726

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*(3), 690–707. https://doi.org/10.1111/ajps.12152

Jensen, C., & Petersen, M. B. (2017). The Deservingness Heuristic and the Politics of Health Care. *American Journal of Political Science*, *61*(1), 68–83. https://doi.org/10.1111/ajps.12251

Johansen, N., Marjanovic, S. V., Kjaer, C. V., Baglini, R. B., & Adler-Nissen, R. (2022). Ridiculing the "tinfoil hats:" Citizen responses to COVID-19 misinformation in the Danish facemask debate on Twitter. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-93

Jørgensen, F., Bor, A., & Petersen, M. B. (2021). Compliance without fear: Individual-level protective behaviour during the first wave of the COVID-19 pandemic. *British Journal of Health Psychology*, *26*(2), 679–696. https://doi.org/10.1111/bjhp.12519

Jost, J. T., & Thompson, E. P. (2000). Group-Based Dominance and Opposition to Equality as Independent Predictors of Self-Esteem, Ethnocentrism, and Social Policy Attitudes among African Americans and European Americans. *Journal of Experimental Social Psychology*, *36*(3), 209–232. https://doi.org/10.1006/jesp.1999.1403

Kalmoe, N. P., & Mason, L. (2022). *Radical American Partisanship: Mapping Violent Hostility, Its Causes, and the Consequences for Democracy*. University of Chicago Press.

Kaye, D. (2021, January 11). *Four Questions About Regulating Online Hate Speech*. Medium. Retrieved April 15, 2021, from https://onezero.medium.com/four-questions-about-online-hate-speech-ae3e0a134472

Keller, D. (2019, September 22). *Facebook Restricts Speech by Popular Demand*. The Atlantic. Retrieved March 10, 2022, from https://www.theatlantic.com/ideas/archive/2019/09/facebook-restricts-free-speech-popular-demand/598462/

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication*. https://doi.org/10.1093/joc/jqab034

Kim, T. (2022). Violent political rhetoric on Twitter. *Political Science Research and Methods*, 1–23. https://doi.org/10.1017/psrm.2022.12

Kingzette, J., Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). How Affective Polarization Undermines Support for Democratic Norms. *Public Opinion Quarterly*, *85*(2), 663–677. https://doi.org/10.1093/poq/nfab029

Klar, S., Krupnikov, Y., & Ryan, J. B. (2018). Affective Polarization or Partisan Disdain? *Public Opinion Quarterly*, *82*(2), 379–390. https://doi.org/10.1093/poq/nfy014

Kowal, M., Coll-Martín, T., Ikizer, G., Rasmussen, J., Eichel, K., Studzińska, A., Koszałkowska, K., Karwowski, M., Najmussaqib, A., Pankowski, D., Lieberoth, A., & Ahmed, O. (2020). Who is the Most Stressed During the COVID-19 Pandemic? Data From 26 Countries and Areas. *Applied Psychology: Health and Well-Being*, *12*(4), 946–966. https://doi.org/10.1111/aphw.12234

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological*

Bulletin, *140*(4), 1073–1137. https://doi.org/10.1037/a0035618

Kowalski, R. M., Limber, S. P., & McCord, A. (2019). A developmental approach to cyberbullying: Prevalence and protective factors. *Aggression and Violent Behavior*, *45*, 20–32. https://doi.org/10.1016/j.avb.2018.02.009

Kteily, N., Bruneau, E., Waytz, A., & Cotterill, S. (2015). The ascent of man: Theoretical and empirical evidence for blatant dehumanization. *Journal of Personality and Social Psychology*, *109*(5), 901. https://doi.org/10.1037/pspp0000048

Kteily, N., Hodson, G., & Bruneau, E. (2016). They see us as less than human: Metadehumanization predicts intergroup conflict via reciprocal dehumanization. *Journal of Personality and Social Psychology*, *110*(3), 343–370. https://doi.org/10.1037/pspa0000044

Kugler, M. B., Cooper, J., & Nosek, B. A. (2010). Group-Based Dominance and Opposition to Equality Correspond to Different Psychological Motives. *Social Justice Research*, *23*(2-3), 117–155. https://doi.org/10.1007/s11211-010-0112-5

Kunst, J. R., Fischer, R., Sidanius, J., & Thomsen, L. (2017). Preferences for group dominance track and mediate the effects of macro-level social inequality and violence across societies. *Proceedings of the National Academy of Sciences*, *114*(21), 5407–5412. https://doi.org/10.1073/pnas.1616572114

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee, N. M. (2018). Fake news, phishing, and fraud: A call for research on digital media literacy education beyond the classroom. *Communication Education*, *67*(4), 460–466. https://doi.org/10.1080/03634523.2018.1503313

Leeper, T. J., & Slothuus, R. (2014). Political parties, motivated reasoning, and public opinion formation. *Political Psychology*, *35*, 129–156. https://doi.org/10.1111/pops.12164

Lelkes, Y., & Westwood, S. J. (2016). The limits of partisan prejudice. *The Journal of Politics*, *79*(2), 485–501. https://doi.org/10.1086/688223

Levendusky, M. S., & Malhotra, N. (2016). (Mis)perceptions of Partisan Polarization in the American Public. *Public Opinion Quarterly*, *80*(S1), 378–391. https://doi.org/10.1093/poq/nfv045

Lieberoth, A., Lin, S.-Y., Stöckli, S., Han, H., Kowal, M., Gelpi, R., Chrona, S., Tran, T. P., Jeftić, A., Rasmussen, J., Cakal, H., & Milfont, T. L.

(2021). Stress and worry in the 2020 coronavirus pandemic: Relationships to trust and compliance with preventive measures across 48 countries in the COVIDiSTRESS global survey. *Royal Society Open Science*, *8*(2). https://doi.org/10.1098/rsos.200589

Maddux, J. E., & Rogers, R. W. (1983). Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of Experimental Social Psychology*, *19*(5), 469–479. https://doi.org/10.1016/0022-1031(83)90023-9

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*, 1–16. https://doi.org/10.1037/xap0000315

Marcus, G. E., Sullivan, J. L., Theiss-Morse, E., & Wood, S. L. (1995). *With malice toward some: How people make civil liberties judgments*. Cambridge University Press. https://doi.org/10.1017/CBO9781139174046

Martherus, J. L., Martinez, A. G., Piff, P. K., & Theodoridis, A. G. (2021). Party Animals? Extreme Partisan Polarization and Dehumanization. *Political Behavior*, *43*(2), 517–540. https://doi.org/10.1007/s11109-019-09559-4

Massaro, T. M., & Stryker, R. (2012, April 18). Freedom of Speech, Liberal Democracy, and Emerging Evidence on Civility and Effective Democratic Engagement. Retrieved January 22, 2023, from https://papers.ssrn.com/abstract=2042171

McGuire, W. J., & Papageorgis, D. (1962). Effectiveness of forewarning in developing resistance to persuasion. *Public Opinion Quarterly*, *26*(1), 24–34.

Mchangama, J., Alkiviadou, N., & Mendiratta, R. (2020). Global Handbook on Hate Speech Laws.

Mernyk, J. S., Pink, S. L., Druckman, J. N., & Willer, R. (2022). Correcting inaccurate metaperceptions reduces Americans' support for partisan violence. *Proceedings of the National Academy of Sciences*, *119*(16). https://doi.org/10.1073/pnas.2116851119

Mill, J. S. (1966). On liberty. In *A selection of his works* (pp. 1–147). Springer.

Moor, L., & Anderson, J. R. (2019). A systematic literature review of the relationship between dark personality traits and antisocial online behaviours. *Personality and Individual Differences*, *144*, 40–55. https://doi.org/10.1016/j.paid.2019.02.027

Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B., & Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the*

National Academy of Sciences, *117*(26), 14864–14872. https://doi.org/10.1073/pnas.2001263117
Mosleh, M., Pennycook, G., & Rand, D. G. (2021). Field Experiments on Social Media. *Current Directions in Psychological Science*, 09637214211054761. https://doi.org/10.1177/09637214211054761
Muddiman, A. (2017). Personal and Public Levels of Political Incivility, 21.
Muddiman, A. (2021, April 2). Conservatives and Incivility. In S. E. Jarvis (Ed.), *Conservative Political Communication* (1st ed., pp. 119–136). Routledge. https://doi.org/10.4324/9781351187237-8
Mukerjee, S., & Yang, T. (2020). Choosing to avoid? A conjoint experimental study to understand selective exposure and avoidance on social media. *Political Communication*, 1–19. https://doi.org/10.1080/10584609.2020.1763531
Munger, K. (2016). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5
Munger, K. (2020). Don't @ me: Experimentally reducing partisan incivility on twitter. *Journal of Experimental Political Science*, 1–15. https://doi.org/10.1017/xps.2020.14
Mutz, D. C. (2015). *In-your-face politics: The consequences of uncivil media*. Princeton University Press.
Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059
OHCHR. (2012). *Rabat Plan of Action*.
Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review*, 1–17. https://doi.org/10.1017/s0003055421000290
Pacilli, M. G., Roccato, M., Pagliaro, S., & Russo, S. (2016). From political opponents to enemies? The role of perceived moral distance in the animalistic dehumanization of the political outgroup. *Group Processes & Intergroup Relations*, *19*(3), 360–373. https://doi.org/10.1177/1368430215590490
Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*(1), 339–367. https://doi.org/10.1146/annurev.psych.60.110707.163607
Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice Reduction: Progress and Challenges. *Annual Review of*

Psychology, *72*(1), 533–560. https://doi.org/10.1146/annurev-psych-071620-030619

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media and Society*, *6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Pasek, M. H., Ankori-Karlinsky, L.-O., Levy-Vene, A., & Moore-Berg, S. L. (2022). Misperceptions about out-partisans' democratic values may erode democracy. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-19616-4

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Petersen, M. B., Bor, A., Jørgensen, F., & Lindholt, M. F. (2021). Transparent communication about negative features of COVID-19 vaccines decreases acceptance but increases trust. *Proceedings of the National Academy of Sciences*, *118*(29), e2024597118. https://doi.org/10.1073/pnas.2024597118

Petersen, M. B., Osmundsen, M., & Arceneaux, K. (2018, September 1). *The "Need for Chaos" and Motivations to Share Hostile Political Rumors* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/6m4ts

Petersen, M. B., Osmundsen, M., & Tooby, J. (2020). The evolutionary psychology of conflict and the functions of falsehood. In D. C. Baker & E. Suhay (Eds.), *The Politics of Truth in Polarized America Concepts, Causes and Correctives*. Oxford University Press. https://doi.org/10.31234/osf.io/kaby9

Pratto, F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741. https://doi.org/10.1037/0022-3514.67.4.741

Pretus, C., Javeed, A., Hughes, D. R., Hackenburg, K., Tsakiris, M., Vilarroya, O., & Bavel, J. J. V. (2022, July 19). The Misleading count: An identity-based intervention to mitigate the spread of partisan misinformation. https://doi.org/10.31234/osf.io/7j26y

Quandt, T. (2018). Dark participation. *Media and Communication*, *6*(4), 36–48. https://doi.org/10.17645/mac.v6i4.1519

Rabinowitz, J. L., Sears, D. O., Sidanius, J., & Krosnick, J. A. (2009). Why Do White Americans Oppose Race-Targeted Policies? Clarifying

the Impact of Symbolic Racism. *Political Psychology*, *30*(5), 805–828. https://doi.org/10.1111/j.1467-9221.2009.00726.x

Rasmussen, S. H. R., & Petersen, M. B. (2022, May 26). From Echo Chambers to Resonance Chambers: How Offline Political Events Enter and Are Amplified In Online Networks. https://doi.org/10.31234/osf.io/vzu4q

Rathje, S., Roozenbeek, J., Traberg, C. S., Bavel, J. J. V., & van der Linden, D. S. (2022, April 2). Letter to the Editors of Psychological Science: Meta-Analysis Reveals that Accuracy Nudges Have Little to No Effect for U.S. Conservatives: Regarding Pennycook et al. (2020). https://doi.org/10.31234/osf.io/945na

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118. https://doi.org/10.1073/pnas.2024292118

Rippetoe, P. A., & Rogers, R. W. (1987). Effects of components of protection-motivation theory on adaptive and maladaptive coping with a health threat. *Journal of personality and social psychology*, *52*(3), 596.

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Bavel, V., & Feuerriegel, S. (Forthcoming). Negativity drives online news consumption. *Nature Human Behaviour*.

Rogers, R. W. (1975). A Protection Motivation Theory of Fear Appeals and Attitude Change1. *The Journal of Psychology*, *91*(1), 93–114. https://doi.org/10.1080/00223980.1975.9915803

Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020). *Psychological Science*, 09567976211024535. https://doi.org/10.1177/09567976211024535

Roozenbeek, J., Traberg, C. S., & Van Der Linden, S. (2022). Technique-based inoculation against real-world misinformation. *Royal Society Open Science*, *9*(5). https://doi.org/10.1098/rsos.211719

Rossini, P. (2019). Disentangling uncivil and intolerant discourse. In R. G. Boatright, T. J. Shaffer, S. Sobieraj, & D. G. Young (Eds.), *A Crisis of Civility? Contemporary Research on Civility, Incivility, and Political Discourse*. Routledge.

Rowe, I. (2015). Deliberation 2.0: Comparing the deliberative quality of online news user comments across platforms. *Journal of Broadcasting & Electronic Media*, *59*(4), 539–555. https://doi.org/10.1080/08838151.2015.1093482

Rudnicki, K., Vandebosch, H., Voué, P., & Poels, K. (2022). Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults.

Behaviour & Information Technology, *0*(0), 1–18. https://doi.org/10.1080/0144929X.2022.2027013

Ruggeri, K., Većkalov, B., Bojanić, L., Andersen, T. L., Ashcroft-Jones, S., Ayacaxli, N., Barea-Arroyo, P., Berge, M. L., Bjørndal, L. D., Bursalıoğlu, A., Bühler, V., Čadek, M., Çetinçelik, M., Clay, G., Cortijos-Bernabeu, A., Damnjanović, K., Dugue, T. M., Esberg, M., Esteban-Serna, C., … Folke, T. (2021). The general fault in our fault lines. *Nature Human Behaviour*, *5*(10), 1369–1380. https://doi.org/10.1038/s41562-021-01092-x

Schmitt, M. T., Branscombe, N. R., & Kappen, D. M. (2003). Attitudes toward group-based inequality: Social dominance or social identity? *British Journal of Social Psychology*, *42*(2), 161–186. https://doi.org/10.1348/014466603322127166

Sellars, A. F. (2016). Defining hate speech. *Berkman Klein Center Research Publication No. 2016-20*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244

Sheeran, P., Aubrey, R., & Kellett, S. (2007). Increasing attendance for psychotherapy: Implementation intentions and the self-regulation of attendance-related negative affect. *Journal of Consulting and Clinical Psychology*, *75*(6), 853–863. https://doi.org/10.1037/0022-006X.75.6.853

Sheeran, P., & Orbell, S. (2000). Using implementation intentions to increase attendance for cervical cancer screening. *Health Psychology*, *19*(3), 283. https://doi.org/10.1037/0278-6133.19.3.283

Sidanius, J., Liu, J. H., Shaw, J. S., & Pratto, F. (1994). Social Dominance Orientation, Hierarchy Attenuators and Hierarchy Enhancers: Social Dominance Theory and the Criminal Justice System. *Journal of Applied Social Psychology*, *24*(4), 338–366. https://doi.org/10.1111/j.1559-1816.1994.tb00586.x

Sidanius, J., & Pratto, F. (2001, February 12). *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. Cambridge University Press.

Sidanius, J., Pratto, F., & Bobo, L. (1996). Racism, conservatism, Affirmative Action, and intellectual sophistication: A matter of principled conservatism or group dominance? *Journal of Personality and Social Psychology*, *70*, 476–490. https://doi.org/10.1037/0022-3514.70.3.476

Siegel, A. (2020). Online hate speech. In N. Persily, J. Tucker, & J. A. Tucker (Eds.). N. Persily (**typeredactor**), *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press. https://doi.org/10.1017/9781108890960

Siegel, A., & Badaan, V. (2020). #No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, 1–19. https://doi.org/10.1017/s0003055420000283

Siegel, A., Nikitinz, E., Barberá, P., Sterling, J., Bethany Pullenk, Bonneau, R., Nagler, J., & Tucker, J. A. (2019). Trumping hate on twitter.

Simonovits, G., Mccoy, J., & Littvay, L. (2022). Democratic Hypocrisy and Out-group Threat: Explaining Citizen Support for Democratic Erosion. *The Journal of Politics*. https://doi.org/10.1086/719009

Skaaning, S.-E., & Krishnarajan, S. (2021). Who Cares About Free Speech? https://futurefreespeech.com/wp-content/uploads/2021/06/Report_Who-cares-about-free-speech_21052021.pdf

Skytte, R. K. (2021). Degrees of Disrespect: How Only Extreme and Rare Incivility Alienates the Base. *The Journal of Politics*. https://doi.org/10.1086/717852

Sniderman, P. M. (2018). Some advances in the design of survey experiments. *Annual Review of Political Science*, *21*(1), 259–275. https://doi.org/10.1146/annurev-polisci-042716-115726

Sniderman, P. M., & Carmines, E. G. (1997). Reaching beyond race. *PS: Political Science & Politics*, *30*(3), 466–471.

Sniderman, P. M., Piazza, T., Tetlock, P. E., & Kendrick, A. (1991). The New Racism. *American Journal of Political Science*, *35*(2), 423–447. https://doi.org/10.2307/2111369

Sniderman, P. M., & Piazza, T. L. (1993). *The scar of race*. Harvard University Press.

Sniderman, P. M., & Tetlock, P. E. (1993). *Prejudice, politics, and the American dilemma*. Stanford University Press.

Sniderman, P. M., Tetlock, P. E., Glaser, J. M., Green, D. P., & Hout, M. (1989). Principled Tolerance and the American Mass Public. *British Journal of Political Science*, *19*(1), 25–45. Retrieved May 21, 2021, from https://www.jstor.org.ez.statsbiblioteket.dk:2048/stable/193786

Stein, J. (2016). *How Trolls Are Ruining the Internet*. Time. Retrieved January 20, 2023, from https://time.com/4457110/internet-trolls/

Stevens, D., Allen, B., Sullivan, J., & Lawrence, E. (2015). Fair's Fair? Principles, Partisanship, and Perceptions of the Fairness of Campaign Rhetoric. *British Journal of Political Science*, *45*(1), 195–213. https://doi.org/10.1017/s0007123413000045

Stouffer, S. A. (1955). *Communism, conformity, and civil liberties: A cross-section of the nation speaks its mind*. Transaction Publishers.

Stromer-Galley, J., & Wichowski, A. (2011). Political discussion online. In *The handbook of internet studies* (pp. 168–187). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444314861.ch8

Suderman, P. (2018). *The Slippery Slope of Regulating Social Media*. Retrieved January 20, 2023, from https://www.nytimes.com/2018/09/11/opinion/the-slippery-slope-of-regulating-social-media.html

Suler, J. (2004). The Online Disinhibition Effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Sullivan, J. L., Piereson, J., & Marcus, G. E. (1982). *Political tolerance and American democracy*. University of Chicago Press.

Svolik, M. (2018, September 3). *When Polarization Trumps Civic Virtue: Partisan Conflict and the Subversion of Democracy by Incumbents* (SSRN Scholarly Paper No. ID 3243470). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3243470

Tappin, B. M., & McKay, R. T. (2019). Moral polarization and out-party hostility in the US political context. *Journal of Social and Political Psychology*, *7*(1), 213–245. https://doi.org/10.5964/jspp.v7i1.1090

Tucker, J. A., Theocharis, Y., Roberts, M. E., & Barberá, P. (2017). From liberation to turmoil: Social media and democracy. *Journal of Democracy*, *28*(4), 46–59. https://doi.org/10.1353/jod.2017.0064

Tyler, T. R. (2006a). Psychological perspectives on legitimacy and legitimation (2005/12/02). *Annu Rev Psychol*, *57*, 375–400. https://doi.org/10.1146/annurev.psych.57.102904.190038

Tyler, T. R. (2006b). *Why People Obey the Law*. Princeton University Press. https://doi.org/10.1515/9781400828609

Uscinski, J. E., Enders, A. M., Seelig, M. I., Klofstad, C. A., Funchion, J. R., Everett, C., Wuchty, S., Premaratne, K., & Murthi, M. N. (2021). American Politics in Two Dimensions: Partisan and Ideological Identities versus Anti-Establishment Orientations. *American Journal of Political Science*, *65*(4), 877–895. https://doi.org/10.1111/ajps.12616

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political Psychology in the Digital (mis)Information age: A Model of News Belief and Sharing. *Social Issues and Policy Review*, *15*(1), 84–113. https://doi.org/10.1111/sipr.12077

Van Bavel, J. J., & Packer, D. J. (2021). *The power of us: Harnessing our shared identities to improve performance, increase cooperation, and promote social harmony*. Little, Brown Spark.

Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C., & Sternisko, A. (2021). How social media shapes polarization. *Trends in Cognitive Sciences*, *25*(11), 913–916. https://doi.org/10.1016/j.tics.2021.07.013

Van Der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*. https://doi.org/10.1038/s41591-022-01713-6

Van Geel, M., Goemans, A., Toprak, F., & Vedder, P. (2017). Which personality traits are related to traditional bullying and cyberbullying? A study with the big five, dark triad and sadism. *Personality and Individual Differences*, *106*, 231–235. https://doi.org/10.1016/j.paid.2016.10.063

Vecchiato, A., & Munger, K. (2022). Validating the Visual Conjoint, with an Application to Candidate Evaluation on Social Media, 58.

Vidgen, B., Margetts, H., & Harris, A. (2019). *How much online abuse is there. a systematic review of evidence for the UK*. The Alan Turing Institute.

Vogels, E. A., Perrin, A., & Anderson, M. (2020). *Most Americans Think Social Media Sites Censor Political Viewpoints*. Pew Research Center.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, *38*(3), 505–520. https://doi.org/10.1016/j.ssresearch.2009.03.004

Walter, N., Brooks, J. J., Saucier, C. J., & Suresh, S. (2020). Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. *Health Communication*, *0*(0), 1–9. https://doi.org/10.1080/10410236.2020.1794553

Westwood, S. J., Grimmer, J., Tyler, M., & Nall, C. (2022). Current research overstates American support for political violence. *Proceedings of the National Academy of Sciences*, *119*(12), e2116870119. https://doi.org/10.1073/pnas.2116870119

Wike, R. (2016). *Americans tolerate offensive speech more than others in world*. Pew Research Center. Retrieved December 7, 2020, from https://www.pewresearch.org/fact-tank/2016/10/12/americans-more-tolerant-of-offensive-speech-than-others-in-the-world/

Wojcieszak, M., Casas, A., Yu, X., Nagler, J., & Tucker, J. A. (2022). Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. *Science Advances*, *8*(39), eabn9418. https://doi.org/10.1126/sciadv.abn9418

Wolchover, N. (2012). *Why Is Everyone on the Internet So Angry?* Scientific American. Retrieved January 13, 2023, from https://www.scientificamerican.com/article/why-is-everyone-on-the-internet-so-angry/

Yu, X., Wojcieszak, M., & Casas, A. (2021). Affective polarization on social media: In-party love among American politicians, greater engagement with out-party hate among ordinary users. https://doi.org/10.31219/osf.io/rhmb9

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X

Zuleta, L., & Burkal, R. (2017). *Hate speech in the public online debate*. Danish Institute for Human Rights.