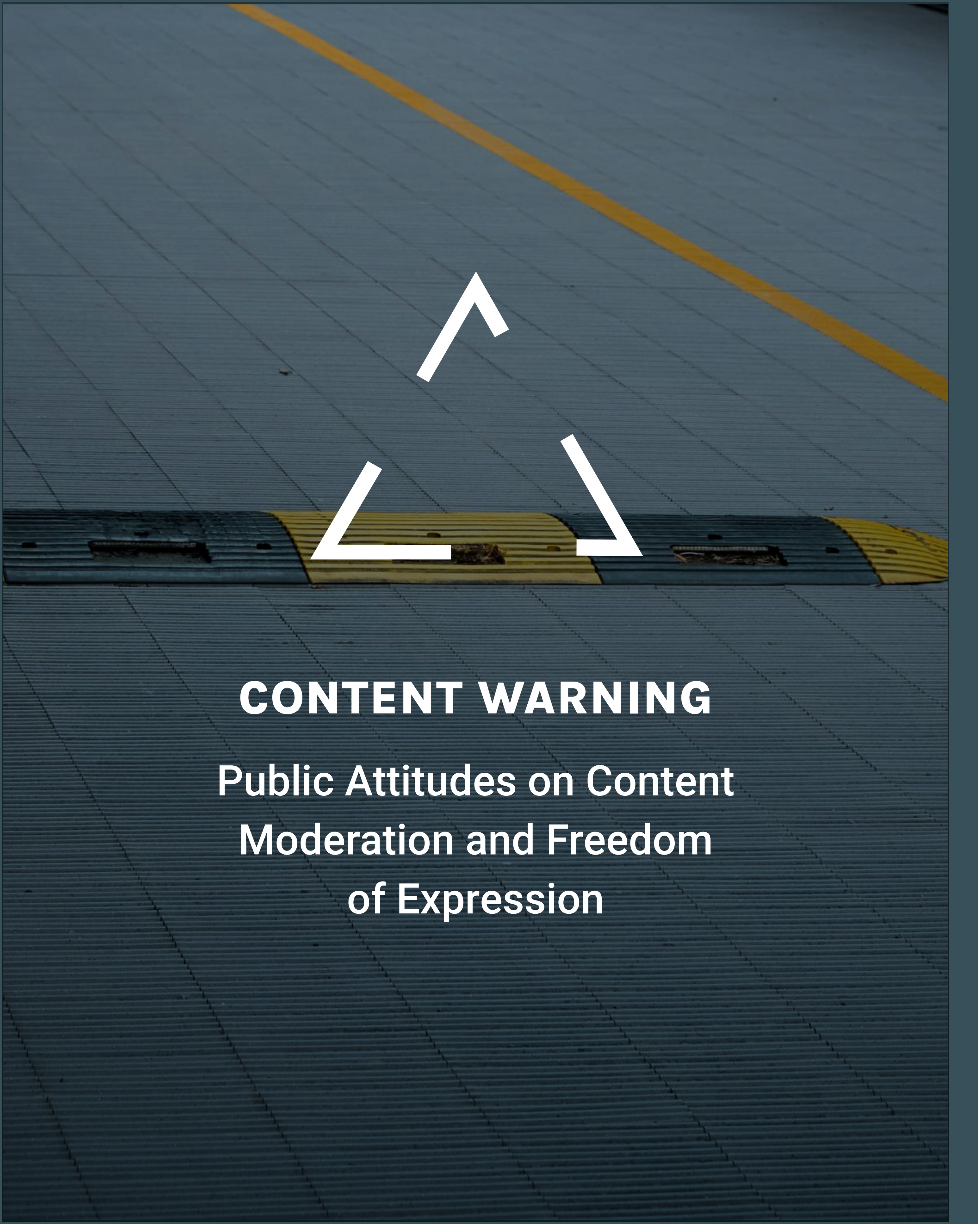


CONTENT WARNING

Public Attitudes on Content Moderation and Freedom of Expression

Yannis Theocharis, Spyros Kosmidis, Jan Zilinsky,
Friederike Quint, Franziska Pradel

CONTENT WARNING

Public Attitudes on Content Moderation and Freedom of Expression

RECOMMENDED CITATION

Theocharis, Y., Kosmidis, S., Zilinsky, J., Quint, F., & Pradel, F. (2025). *Content Warning: Public Attitudes on Content Moderation and Freedom of Expression*. Content Moderation Lab at TUM Think Tank. https://doi.org/10.17605/OSF.IO/F56BH

Chair of Digital Governance
Munich School of Politics and Public Policy
Technical University of Munich

CONTENT
MODERATION LAB

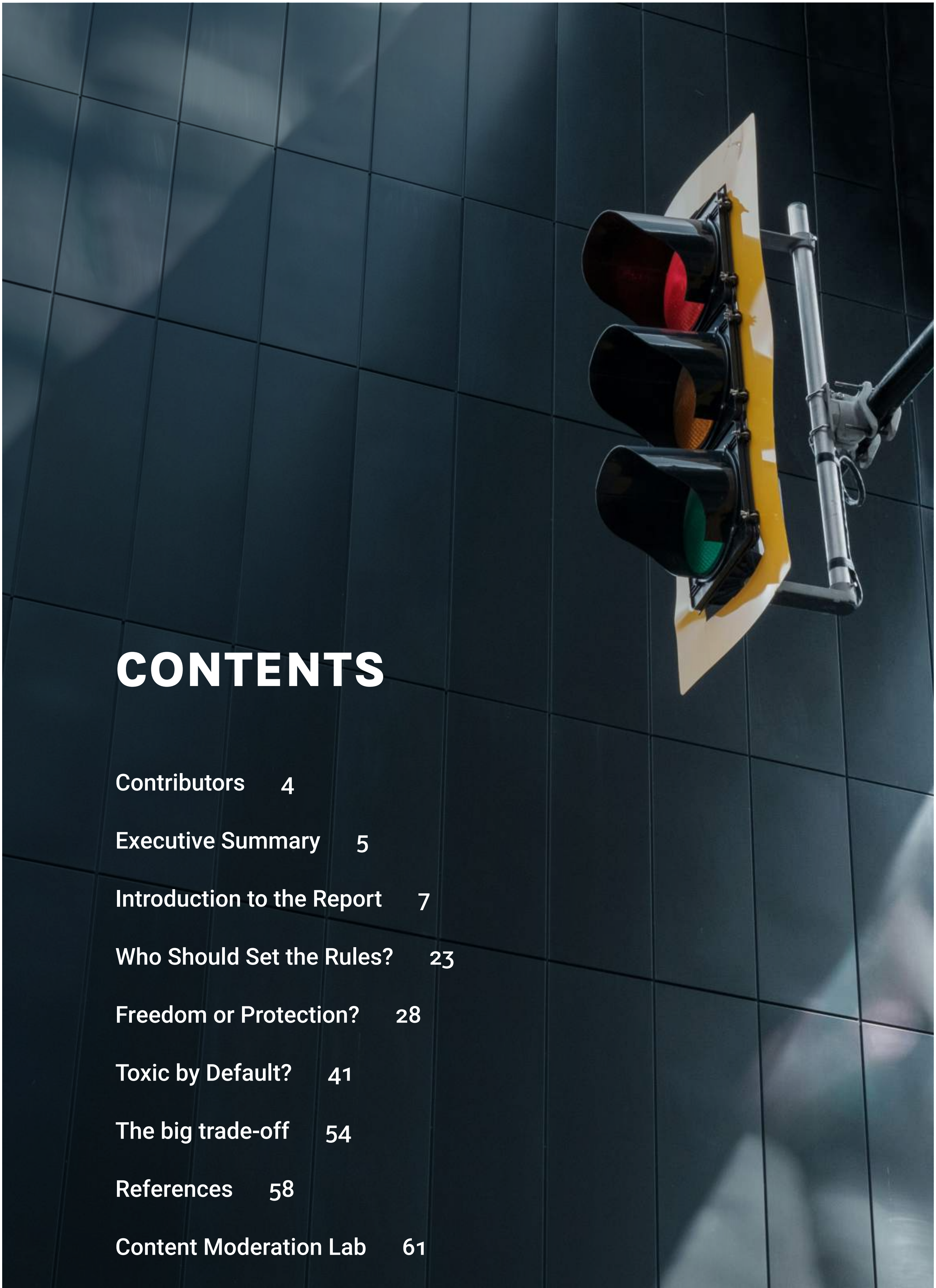
TUM THINK TANK

Technical
University
of Munich

TUM

UNIVERSITY OF
OXFORD

DPIR
DEPARTMENT OF POLITICS &
INTERNATIONAL RELATIONS



CONTENTS

CONTRIBUTORS

Yannis Theocharis is Professor and Chair of Digital Governance at Technical University of Munich, and Co-Principal Investigator of the Content Moderation Lab at the TUM Think Tank. His field of research is political behavior with a focus on how digital media impact political participation, communication, governance, and uncivil behavior in online spaces. His work deploys survey, experimental, and computational methods to understand how social media changes the political information environment, empowers users by enabling new forms of political participation, but also demobilizes, marginalizes, and polarizes the public.

Spyros Kosmidis is Associate Professor in the Department of Politics and International Relations at the University of Oxford and the Director of the Oxford Q-Step Centre, Director of the Oxford Spring School in Advanced Methods, and Director of the MPhil in Comparative Government at the University of Oxford. Further, he is Co-Principal Investigator of the Content Moderation Lab at the TUM Think Tank at Technical University of Munich. His work focuses on public opinion, political behavior, social attitudes, and party competition.

Jan Zilinsky is a post-doctoral researcher at the Chair of Digital Governance at the Technical University of Munich. He is also a research affiliate with the NYU Center for Social Media and Politics and with the Institute of Experimental Psychology at the Slovak Academy of Sciences.

Friederike Quint is a doctoral candidate and research associate at the Chair of Digital Governance at the Technical University of Munich and member of the Content Moderation Lab. Her research focuses on the transparency and impacts of online content moderation, platform governance, public perceptions and attitudes of social media, survey experiments and computational social science.

Franziska Pradel is a post-doctoral researcher at the Chair of Digital Governance at the Technical University of Munich. Her research interests include online political communication, especially investigating biases on online platforms and their effects on political attitudes, computational social science, and experiments.

EXECUTIVE SUMMARY

This report investigates public attitudes toward content moderation, balancing freedom of speech with harm prevention, and the prevalence of toxic behavior in online spaces. Drawing on representative survey data from ten countries, the findings reveal a detailed picture of how citizens perceive the role of social media platforms, governments, and independent organizations in moderating content, and how they reconcile competing priorities for free expression and safety.

The report shows that across nations, there is no universal consensus on who should bear the responsibility for maintaining a healthy online environment. While platforms are often viewed as best positioned to combat harmful speech, the survey results highlight variations in public expectations about their roles. These findings suggest a need for greater clarity and accountability in content moderation policies, reflecting diverse cultural and political contexts.

The report reveals a complex relationship between support for free speech and concerns about harm prevention. While countries like Sweden, Greece, the US, and Germany lean toward protecting free speech, others such as South Africa, Brazil, and France prioritize harm prevention. Despite these differences, most respondents favor a balanced approach, challenging the notion that unrestricted speech inherently upholds democratic values. Instead, our findings highlight the public's growing awareness that unchecked harmful content can marginalize voices and undermine democratic principles.

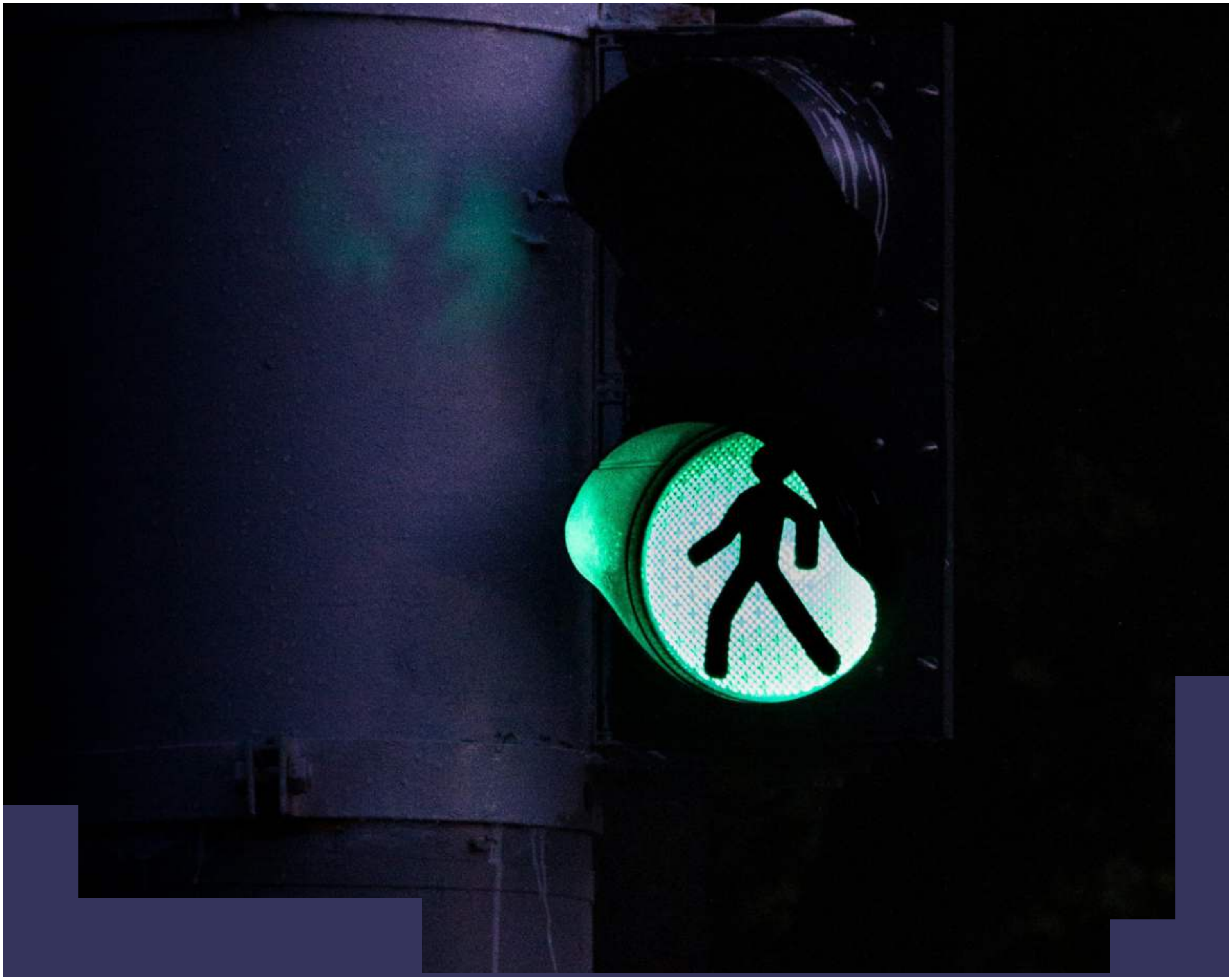
A troubling trend emerges in the normalization of toxic behavior both online and offline. Many respondents perceive incivility, hate, and discrimination as inevitable aspects of modern social engagement, reflecting a deep sense of resignation about the capacity of platforms to address these issues. This perception aligns with respondents' lived experiences online, with significant numbers reporting verbal abuse, threats, and discrimination rooted in sensitive identity factors such as race, gender, political views, and sexual orientation.

When presented with hypothetical trade-offs between freedom of expression and content moderation, respondents overwhelmingly favored moderation to reduce harmful content. Regarding hateful content, most respondents across countries, including the US, preferred platforms that would actively remove intolerant, uncivil, or discriminatory posts. The demand for moderation was even stronger concerning misinformation, with respondents prioritizing platforms free of fake news over unrestrained freedom to post. Countries like France, Germany, and Brazil showed the highest preference for moderation, reflecting global concerns about the impact of harmful content.

The study underscores the importance of understanding free speech beyond abstract principles. By presenting respondents with tangible scenarios, it highlights the complex and often contradictory attitudes people hold about free speech and prevention from harm. This approach sheds light on the boundaries respondents place on free expression, shaped by cultural norms and national contexts.

The ongoing debate over content moderation, fueled by policy changes at major platforms like Meta and X (formerly Twitter), has significant implications for public trust and engagement. While some platforms have adopted a laissez-faire approach, our findings suggest this does not align with public preferences. Most users want platforms to actively reduce hate speech and misinformation, viewing moderation as essential to healthy online environments. The experiences of platforms like X, where engagement and profitability have suffered under minimal moderation, further underscore the risks of prioritizing unrestricted speech over user safety.

The findings highlight a global consensus on the importance of moderation to protect individuals and maintain healthy discourse in digital spaces. While national traditions and cultural norms shape specific preferences, the overarching message is that citizens value freedom of speech but recognize the need for safeguards against harmful content. Policymakers and platform leaders must consider these public attitudes to ensure digital spaces remain both safe and inclusive while respecting the principles of free expression.



Introduction to the Report

Background and Key Terms

Social media platforms have become integral to daily life, shaping how we connect, communicate, and consume information. Studies indicate that a large portion of society relies on social media for news, enabling interactions with diverse groups and facilitating exchanges with individuals and organizations that would be unlikely in offline settings (Pew Research Center, 2021; Newman et al., 2021; Ellison & Vitak, 2015). These platforms have revolutionized the democratic process by encouraging political participation and diversifying news audiences and consumption patterns (Boulianne, 2018; Fletcher & Nielsen, 2017).

Yet, the very traits that make social media so important for politics, and especially for the articulation and consumption of political speech, also make it contentious (Tucker et al., 2017). Initially seen as spaces that foster deliberation and can open users to diverse—especially contrarian—perspectives, there is now evidence that these platforms enable worrying levels of hate speech, mis- and disinformation, and societal division (Lorenz-Spreen et al., 2023; Rathje et al., 2021). Given that all these detrimental phenomena have been repeatedly linked to adverse psychological and physiological effects influencing aspects of human behavior that go well beyond intergroup dynamics—including the erosion of anti-discriminatory norms and desensitization to harmful speech (Bilewicz & Soral, 2020; Relia et al., 2019; Pluta et al., 2023, Dreißigacker et al., 2024)—the regulation of speech on social media has become an urgent matter of discussion and a salient topic in public discourse.

The very traits that make social media so important for politics and political speech also make it contentious.

Amid these discussions, the focus on **content moderation** has become increasingly central, as it encompasses the governance mechanisms designed to structure user interactions and prevent harmful behaviors. Content moderation refers to “the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse.”

(Grimmelmann, 2015:6). It involves design decisions that structure user interactions, and practices like the suspension of individual accounts (Gorwa et al., 2019). Platforms implement content moderation through several approaches. Automated detection systems use algorithms and machine learning to scan content for harmful language, misinformation, or graphic material. Social media platforms publicly state that their AI engines are highly effective at finding and deleting such content. But as the case of Facebook has shown, while efficient for handling large volumes of data, these systems often struggle to account for contextual nuance, leading to serious errors in enforcement (Wired, 2021, see also Meta, 2025). In addition to automated approaches, community reporting mechanisms empower users to flag inappropriate content for review, leveraging the platform's user base to identify issues.

Content moderation guidelines also play a critical role. Not only are they essential for setting standards and ensuring consistency in how platforms enforce rules around acceptable behavior and content, but they can help users make informed decisions on what speech is—and is not—admissible (e.g. Singhal et al. 2023). These guidelines have been partly influenced by civil society initiatives such as the Santa Clara Principles, which emphasize transparency, accountability, and the protection of marginalized voices in content moderation practices (even though the effectiveness of multistakeholder governance, as exemplified by partnerships between civil society organizations and corporations, has been subject to debate—Dvoskin, 2024). Human moderators also remain an essential part of the content moderation process. They are tasked with reviewing flagged content and making judgment calls in ambiguous cases, though this role can have severe psychological consequences due to exposure to disturbing material (Roberts, 2019). Finally, content filters and emergency changes to algorithmic feeds, such as Facebook's "break glass" measures, are deployed to block or limit the spread of harmful material in response to specific events like elections—though very little is known about their exact range and effects (Tech Policy Press, 2024).

Theoretically, all these measures aim to balance the need for open expression with the imperative to mitigate abuse and harm in online spaces. They are also implemented voluntarily by platforms, which in the United States (US), for example, are not legally obligated to moderate content under the provisions of Section 230 of the Communications Decency Act (CDA) (Whitehouse, 2023). But while these measures have been largely welcomed by some, many see them as an effort by platforms to police speech and enforce censorship—a criticism coming from both the left and right of the political spectrum, largely due to distrust towards platforms (for example, three out of four Americans feel it is very likely that social media intentionally censor political viewpoints—Pew Research Center, 2020). **Freedom of speech** is broadly understood here as the right to express opinions and ideas without fear of censorship, government retaliation, or societal suppression (according to Article 19 of the Universal Declaration of Human Rights "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.").

Freedom of speech: the right to express opinions and ideas without fear of censorship, government retaliation, or societal suppression.

On one hand, conservative critics argue that content moderation often targets conservative voices, limiting free expression under the guise of combating misinformation (Pew Research Center, 2020). On the other hand, progressive voices have expressed concerns about the suppression of marginalized voices and the potential for content moderation to silence important critiques of societal power structures (Brennan Center for Justice, 2021). Ultimately, these criticisms reflect deep-seated concerns over how platforms enforce their rules, with many arguing that the approaches are often driven more by political and economic pressures than by genuine efforts to maintain fairness and transparency.

The Responsibility for Content Moderation

As the above overview highlights, there are four key actors involved in content moderation—**governments**, **civil society**, **platforms**, and **citizens**—with individual citizens playing a limited role, their primary function confined to reporting harmful content via platform reporting mechanisms, with only a small fraction of users actively engaging in this process. While these actors are all engaged in the process on content moderation in different ways and have different jurisdictions (e.g. civil society organizations, for example, do not moderate content themselves, but they can act as shapers of content moderation by developing good practices or by offering reporting portals outside the platform-embedded reporting mechanisms), they are all operating at the same time and face distinct challenges.

Platforms face significant trade-offs as they attempt to balance competing priorities. On the one hand, they are expected to maintain environments that support free expression and public discourse while preventing the proliferation of harmful content such as hate speech, disinformation, and incitement to violence (Howard, 2021; Bollinger & Stone, 2022). On the other hand, their commercial interests—such as maximizing user engagement and minimizing operational costs—not only often conflict with the substantial resources and infrastructure required for robust and equitable content moderation (Gillespie, 2018), but they also arguably encourage leaving harmful content on the platform to reap the benefits of the attention it often garners (New York Times, 2023). These decisions profoundly influence public discourse, determining which voices are amplified and which are suppressed, often with disproportionate impacts on marginalized groups (Baribi-Bartov et al., 2024).

Determining which voices are amplified and which are suppressed, often has disproportionate impacts on marginalized groups.

Governments also face complex challenges and trade-offs in this domain. As the primary stewards of public safety and rights, they are tasked with crafting policies that curb harmful online behavior while preserving freedom of expression, a cornerstone of democratic values. Striking this balance is particularly fraught, as excessive regulation risks veering into censorship, stifling dissent, or empowering authoritarian control over speech. Conversely, insufficient oversight can leave harmful content unchecked, with real-world consequences such as political radicalization or harm to vulnerable communities (Müller & Schwarz, 2020; Williams et al ., 2020). Governments must also contend with the global nature of social media, where platforms operate across jurisdictions with varying legal frameworks and cultural norms, complicating enforcement and regulatory coherence. Despite these challenges, there are instances where, with the necessary political will, governments have demonstrated that they have the upper hand and can assert decisive control over platforms, as seen in various cases of speech moderation across both democratic and non-democratic contexts (York, 2022).

Civil society, which encompasses advocacy groups, nonprofit organizations, academic experts, and citizen movements, plays a critical role in holding both platforms and governments accountable. These organizations fulfill a variety of functions, ranging from overseeing content moderation and monitoring the proliferation of hateful content to proposing best practices and offering alternative mechanisms for reporting illegal online activity. However, civil society faces significant challenges of its own. While its actors advocate for greater transparency, inclusivity, and fairness in content moderation, they often lack the resources or unified voice necessary to influence major platforms or global policy debates effectively. Moreover, there is no assurance that their efforts will be adopted by social media companies (Dvoskin, 2024; York, 2022). Success is possible, as demonstrated by the partial adoption of the Santa Clara Principles by some companies, driven by growing demands from users and policymakers. Furthermore, tensions can emerge within civil society itself, as organizations representing divergent values may hold conflicting views on the acceptable boundaries of free expression. For instance, some groups prioritize combating hate speech and safeguarding vulnerable communities, while others caution against the potential chilling effects on legitimate activism and countercultural expression.

Citizens play a multifaceted yet underutilized role in content moderation.

Lastly, citizens play a multifaceted yet underutilized role in content moderation. Primarily, they can report harmful content through platform-provided reporting mechanisms, a process that depends heavily on user awareness and willingness to engage. Beyond reporting, citizens hold significant indirect power by influencing broader systems that govern content moderation. They can advocate for changes by pressuring elected representatives to craft policies that ensure transparency and fairness in moderation practices. Additionally, they wield economic and social influence by choosing to abandon platforms that fail to uphold their standards, thereby sending a clear message to companies about user priorities.[1] Individual citizens may also contribute to shaping public discourse by raising awareness about harmful practices or moderation failures, fostering collective accountability. Despite these opportunities, the overall impact of citizen involvement is limited by uneven participation and the often-opaque nature of content moderation systems. To maximize their influence, citizens need accessible tools, education on the implications of harmful content, and channels to meaningfully engage with policymakers and platforms alike.

Taken together, these actors navigate a web of competing priorities and shared responsibilities. Platforms, governments, civil society, and citizens must grapple with difficult trade-offs between freedom and safety, transparency and efficiency, and regulation and innovation. The interplay between these stakeholders will shape the future of online discourse, influencing the contours of democracy and the public sphere in the digital age.

[1] Following Elon Musk's acquisition of Twitter and his decision to dismiss the majority of its content moderation team, millions of users gradually abandoned the platform—a trend that accelerated significantly after the 2024 US election—inflicting severe damage to X's brand and value (even though the fleeing user base was not the sole factor responsible for this decimating decline) (Mashable, 2024).

What is This Report About?

This report is about the critical but underexplored role of the public—social media's end users—in the ecosystem of content moderation. It examines the unique trade-offs faced by individual citizens as they navigate the benefits and risks of digital communication, including the tension between safeguarding freedom of expression and addressing harms like hate speech, misinformation, and online harassment. While freedom of expression empowers individuals to engage in public discourse, participate in activism, and amplify marginalized voices, the openness of digital platforms also leaves users vulnerable to harmful content with real-world consequences. Those who are frequent targets of such content may face additional hurdles, which can include increased self-censorship and a diminished presence in public discourse.

We focus on the public because, for users, this trade-off is especially challenging given that the lines between legitimate expression and harmful content are often blurry and deeply subjective. What one person views as constructive criticism or satire by a comedian, another might perceive as harassment or abuse. Citizens must navigate these gray areas while relying on platforms, governments, and civil society to mediate boundaries that are often inconsistently applied or poorly understood. Complicating matters further, individuals' personal values and experiences shape their tolerance for harmful content. Some prioritize an unrestricted digital space, valuing the ability to encounter all ideas, no matter how uncomfortable. Others seek greater safeguards against speech that threatens their dignity, safety, or inclusion in public discourse.

While this subjectivity makes understanding citizens' views challenging, understanding public opinion on these issues is critical for several reasons.

- First, citizens' perceptions and demands often drive changes in platform policies, as seen in cases where public pressure has led to stricter content moderation or new safety features (Gillespie, 2018). Social media platforms are products designed for their users, and citizen feedback can serve as a catalyst for reforms aimed at addressing societal concerns.

- Second, users' collective behavior shapes the algorithms that govern content visibility. Public awareness and criticism of platforms' (lack of) action, particularly when it disproportionately harms marginalized groups, underscore the need for greater transparency and inclusivity in moderation processes. Without understanding how citizens perceive these challenges, efforts to regulate or reform platforms' risk being disconnected from the concerns of those most affected.

- Third, public opinion offers a lens through which the tension between freedom of speech and harm mitigation can be understood in a nuanced way. Individual users often have conflicting views about what content should be allowed or restricted, reflecting diverse cultural, political, and ethical values. By examining these perspectives, policymakers, platform designers, and advocates can better navigate the complex landscape of social media communication, ensuring that any interventions align with societal priorities and democratic principles. Given the global influence of social media, understanding these dynamics is not just desirable—it is essential for shaping a fairer, more transparent and more accountable digital ecosystem.

It is important to note that the trade-off we are presenting here is not static but evolves with societal norms, political climates, and technological advancements. For instance, the rise of algorithm-driven amplification means that citizens are not merely passive recipients of speech but are subjected to curated experiences that prioritize engagement—often at the cost of amplifying polarizing or harmful content. This amplifies the stakes for the public, as the balance between freedom of speech and harm is influenced by opaque systems beyond individual control. **Ultimately, understanding how citizens perceive and navigate these trade-offs is vital because their collective choices and demands shape—directly or indirectly—the platforms and policies that govern the digital age.**

A gap in our knowledge and the need for international comparison

The report was produced after the realization that, while policymakers, platforms, and civil society actors often dominate the conversation, the attitudes of citizens—those most directly impacted by the dynamics of online discourse—remain largely unexplored from an empirical perspective. With few exceptions, such as limited studies conducted in the US by institutes such as the Pew Research Center and the Cato Institute, there is little insight into what citizens actually think about these issues. For instance, do they trust social media platforms and the content they encounter there? Who do they believe should be responsible for moderating content—platforms, governments, independent bodies, or themselves by using models such as those of Wikipedia and Reddit? Most critically, how do they perceive the trade-off between freedom of speech and protection from harm in the digital space? These fundamental questions are essential for shaping policies and practices that resonate with public values, yet they remain unanswered on a global scale.

This gap in public opinion research is especially evident in countries outside of the US and Europe, although even within Europe, perspectives can vary dramatically. The global nature of social media demands a broader understanding of attitudes in diverse cultural, political, and legal contexts. How do citizens in Greece, with its legacy of political instability and concerns about populism, view the regulation of speech compared to those in Germany, a country with stringent laws against hate speech rooted in its historical experience with authoritarianism? In the US, where social media is a political mobilization tool in a deeply polarized society, or in Brazil, where platforms play a central role in contentious elections, public opinion on content moderation may reflect unique national challenges. South Africa's history of apartheid and ongoing struggles with inequality might shape its citizens' views on the role of platforms in addressing hate speech and marginalization. Similarly, in Australia, strict defamation laws and the recent enactment of the Online Safety Act—as well as a social media ban for those under 16 years old—highlight a national focus on combating online harm, shaping public expectations for platform accountability while sparking debates on balancing safety with free expression. Without comparative insights across these varied contexts, platforms, regulators and civil society organizations risk proposing or implementing one-size-fits-all solutions that fail to account for the nuances of global public opinion.

The global nature of social media demands a broader understanding of attitudes in diverse cultural, political, and legal contexts.

These gaps are particularly significant given the stark differences in how governments approach social media regulation across countries. For example, France has proposed drastic measures, with President Emmanuel Macron suggesting platform shutdowns during riots to curb unrest (Washington Post, 2023). The United Kingdom (UK) has implemented stringent measures too, including imprisoning individuals who incited violence via social media during the 2024 riots (BBC, 2024). In contrast, the US' strong emphasis on free speech has led to a more laissez-faire approach, even in extreme cases like the "Unite the Right" rally (New York Times, 2017), where neo-Nazi groups used social media to incite violence that resulted in fatalities and injuries. Meanwhile, countries like Sweden and Australia, known for their strong democratic traditions, grapple with balancing transparency and accountability in moderation. Slovakia's emerging digital policies and Brazil's contentious elections and the controversy around the Supreme Court's decision to temporarily ban X further illustrate the diversity of national approaches.

Understanding public opinion on these issues across such varied contexts is critical. Citizens' views influence the legitimacy of platform policies and government regulations, particularly as they relate to contentious trade-offs between free expression and harm mitigation. Comparative data can illuminate how cultural, historical, and institutional factors shape attitudes, offering insights into the values and priorities of different societies. Furthermore, addressing these gaps can foster a more inclusive conversation about the future of online discourse, ensuring that the rules governing digital spaces reflect the perspectives and experiences of diverse populations.

This report seeks to address these gaps by exploring public opinion across Greece, Germany, the UK, the US, Brazil, South Africa, Slovakia, France, Australia, and Sweden, providing a comprehensive and comparative view of one of the most pressing debates of our time.

The results we present in this report are grounded in rigorous academic research and will contribute to scholarly discourse through published papers. We took the decision to also publish our findings in the form of this report because we believe that they hold significant value for the broader public. Citizens, policymakers, and social media platforms alike are directly affected by the dynamics of content moderation, yet public understanding of these processes remains limited. By sharing our findings beyond academic circles, we aim to bridge this gap, fostering informed public dialogue and empowering individuals to engage meaningfully in discussions about the future of online discourse. Moreover, the decisions made around content moderation impact democratic values, public safety, and freedom of expression—issues that resonate deeply with society at large. Making these insights accessible ensures that the knowledge generated by academia contributes to actionable change in real-world contexts.

Public Opinion on Content Moderation and Free Expression: Survey Coverage Across 10 Countries

Sweden
United Kingdom
Germany
Slovakia
France
Greece
United States
Brazil
Australia
South Africa

Created with Datawrapper

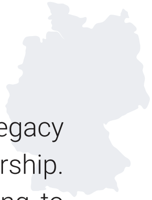
Participating countries

This report is based on representative survey data from 10 countries, encompassing diverse cultural backgrounds, regulatory systems, and democratic traditions. The surveyed countries include Greece, the United States, the United Kingdom, Germany, France, Brazil, Slovakia, Sweden, Australia, and South Africa, offering a broad perspective on public opinion regarding content moderation and free expression.

The countries we study

UNITED STATES

The US was selected for this study due to its pivotal role in shaping global social media and its distinctive approach to free speech. Home to many of the largest social media platforms, the US serves as a critical locus for debates on content moderation. Underpinned by the First Amendment, US law establishes a high threshold for restricting speech, guided by the "imminent lawless action" test established in Brandenburg v. Ohio (1969). This standard evolved from earlier rulings, including the "clear and present danger" test articulated in cases like Schenck v. United States (1919) and later narrowed in Brandenburg to allow state intervention only when speech incites imminent unlawful action and is likely to result in such action. Justice Louis Brandeis's influential concurring opinion in Whitney v. California (1927), although predating Brandenburg, emphasized the value of counterspeech as a remedy for harmful ideas, advocating for more speech rather than enforced silence. At the same time, Section 230 of the Communications Decency Act grants platforms significant autonomy in moderating content, enabling them to become global leaders in social media innovation. However, debates over content moderation reflect deep societal polarization, with conservative groups decrying alleged censorship, progressive voices calling for stronger protections against hate speech and misinformation, and widespread distrust in platforms' ability to effectively police speech. These dynamics, shaped by legal principles and cultural values, make the US an essential case for understanding public perceptions of content moderation and freedom of expression.

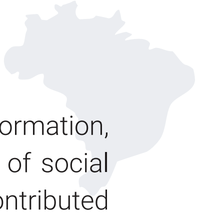


GERMANY

Germany presents a unique case for examining content moderation, shaped by a historical legacy that underscores both the dangers of unchecked speech and the perils of excessive censorship. The Nazi era demonstrated how propaganda and hate speech could fuel atrocities, leading to strong societal and legal commitments to curbing harmful ideas. At the same time, the surveillance culture of the Stasi in East Germany highlighted the risks of state overreach and suppression of dissent. These dual legacies foster a dual imperative in Germany's approach to speech: the need to protect individuals and democracy from harmful content while maintaining vigilance against overly restrictive measures that could undermine free expression. The Network Enforcement Act (NetzDG—now mostly replaced by the EU's Digital Services Act), passed in 2017, exemplifies this balance. The law compels social media platforms to remove illegal content, such as hate speech and Holocaust denial, within strict time frames or face substantial fines. While NetzDG has been praised for addressing the spread of harmful material, it has also sparked debates about overreach, with critics warning against the privatization of censorship and potential threats to free speech. Organizations like Reporters Without Borders, among many others, repeatedly criticized the law, citing the fact that authoritarian governments such as that of Russia adopted a nearly identical regulation shortly thereafter. These concerns are amplified by memories of past abuses of speech regulation, creating a tension between the necessity of moderation and fears of stifling dissent. Public opinion in Germany mirrors these complexities, with strong support for curbing hate speech but a wariness of granting excessive power to platforms or the state. This historical and cultural backdrop, combined with Germany's regulatory leadership through NetzDG, makes it an essential case for understanding how societies navigate the trade-offs between protecting against harm and safeguarding freedom of expression.

AUSTRALIA

Australia's Online Safety Act of 2021 marked a significant step in regulating digital spaces to protect citizens, particularly vulnerable groups, from online harm. The law empowers the eSafety Commissioner to remove harmful content, including cyberbullying and abusive material, and holds platforms accountable through fines if they fail to comply. A recent, vital component of the legislation is the approval of a ban on social media access for users under 16 (the Online Safety Amendment (Social Media Minimum Age) Act of 2024), which is aimed at reducing exposure to harmful content and bullying. Additionally, the act will impose monetary penalties on social media companies that do not take reasonable measures to prevent minors from creating accounts. While praised for addressing online safety concerns, the law has led to reactions by social media companies and has sparked debates about balancing protection with freedom of expression. An opinion poll by YouGov (2024), however, suggests that the majority of Australians are in favour of the age limit and a vast majority of them support the introduction of stronger penalties for social media companies that fail to comply with Australian laws.



BRAZIL

Brazil's social media landscape has been significantly influenced by the role of misinformation, particularly during former President Jair Bolsonaro's administration. His extensive use of social media platforms to disseminate false information about elections and public health contributed to political polarization and unrest. This environment facilitated the organization of the January 2023 riots, where social media played a pivotal role in mobilizing participants. In response to these challenges, Brazil's Supreme Court later took decisive action against the platform X (formerly Twitter) for its role in spreading misinformation. In August 2024, Judge Alexandre de Moraes ordered the suspension of X's operations within Brazil, citing the platform's failure to follow local regulations aimed at combating disinformation and hate speech. This unprecedented move affected approximately 22 million Brazilian users and sparked a global debate on the balance between platform responsibility, freedom of expression, and state overreach. The suspension was lifted in October 2024 after X agreed to comply with the court's directives, including appointing a legal representative in Brazil and paying fines exceeding $5 million.

FRANCE

France represents a critical case study in the evolving landscape of social media content moderation, particularly given its recent assertive stance on platform regulation. With events like the Charlie Hebdo attacks, sparking an intense national debate about the balance between free expression and security as early as 2015, the country has taken increasingly stringent measures to combat harmful online content, as evidenced by the passage of legislation requiring social media companies to remove certain content within one hour. This regulatory approach has been further emphasized by high-profile actions, such as the prosecution of Telegram's CEO for alleged failure to moderate criminal content effectively. The French government, led by President Macron, has also publicly confronted social media platforms over their role in amplifying social unrest, particularly during periods of civil disorder. The French case is thus particularly relevant as it exemplifies the tensions between maintaining platform accountability and preserving free expression in a major European democracy.

GREECE

Greece's approach to social media and speech moderation is shaped by its history, including the military dictatorship of 1967–1974, which saw widespread censorship, and the Nazi occupation, marked by informants and repression. These experiences fostered a wariness of state overreach and a strong commitment to free expression in its democratic era. Yet, recently, a proposal to restrict social media use for minors under 16 gained momentum following public outcry over a brutal incident where a 14-year-old was beaten by peers in an attack organized via social media. This event intensified concerns about the harmful influence of online platforms on youth, fueling debates about the need for stronger protections. However, critics warn that such measures risk encroaching on personal freedoms, reviving historical fears of excessive state control. This ongoing discourse underscores Greece's challenge to balance safeguarding vulnerable populations with preserving free speech in the face of modern digital complexities.

SLOVAKIA

Slovakia represents a compelling case for examining content moderation attitudes, as it exemplifies the complex tensions between protecting free speech and preventing online harm in an emerging digital democracy. The country has recently faced significant challenges related to moderating social media content, especially amid political violence and extremism. While some Slovak politicians have traditionally advocated against content moderation, arguing it could compromise free speech, recent events have highlighted the real-world consequences of unmoderated online discourse. This was starkly illustrated in 2024 when Meta intervened by deleting Facebook accounts following an assassination attempt on former Prime Minister Robert Fico, leading to heated debates about the responsibility of both platforms and political elites. Adding to this complexity, the Slovak government's subsequent decision to prosecute citizens who praised the assassination attempt online marked a significant shift in their approach to "content moderation," creating a notable paradox given their previous stance against intervention.

SOUTH AFRICA

As a country that transitioned from apartheid to democracy, South Africa faces ongoing challenges with online misinformation and cyberbullying, often intensified by deep socioeconomic inequalities (Wassermann 2020). Balancing free speech with the prevention from harm, the country generally leans toward prioritizing the latter. While Section 16 of the Constitution protects freedom of expression, it excludes incitement to violence and hate speech. Laws such as the Protection of Personal Information Act and the Hate Speech Bill further emphasize harm prevention, although this sometimes limits broader free speech. The digital divide in South Africa remains significant, with disparities in Internet infrastructure and affordability affecting access to social media. Since its implementation in 2021, the Protection of Personal Information Act has established a foundation for privacy rights by mandating consent-based data processing and granting individuals the ability to access, correct, and delete their personal data. In addition, the Cybercrimes Act, which has been in effect since 2021, criminalizes harmful data messages and cyber offenses, illustrating the country's proactive approach to safeguarding online interactions. The United Nations has issued a warning about the imminent risk of xenophobic violence in South Africa, highlighting the need for better content moderation on digital platforms (Legal Resources Centre, 2020). Given the country's linguistic diversity and cultural nuances, understanding these factors is crucial for effective moderation. These issues, along with debates around AI ethics, digital taxation, and the regulation of electronic communications, make South Africa an important case study for examining public attitudes toward free speech, privacy, and social media governance in a context marked by socioeconomic disparities.

SWEDEN

Sweden has long been recognized for its strong democratic principles and commitment to free speech, with laws such as the Freedom of the Press Act and the Fundamental Law on Freedom of Expression ensuring transparency and press freedoms. However, in recent years, the country has encountered significant challenges regarding online discourse. Concerns over hate speech, disinformation, and polarization, particularly on social media platforms, have become increasingly prevalent. Notable incidents involving harassment of journalists, politicians, and minority groups, along with actions like the burning of religious texts, have raised important questions about how to balance free speech with the need for safer digital spaces (Disinfo.eu, 2023). Moreover, Sweden has become a target of foreign disinformation campaigns, revealing its vulnerability to online manipulation. In response, the Swedish government has taken proactive steps to promote digital literacy and tackle harmful online content, making the country a particularly relevant case for examining public attitudes toward content moderation and the evolving complexities of free speech in the digital age.

UNITED KINGDOM

In the UK, the Online Safety Act 2023 introduced comprehensive regulations requiring social media companies and search services to implement systems that mitigate illegal activities and swiftly remove illegal content. The Act places particular emphasis on safeguarding children, mandating platforms to prevent access to harmful and age-inappropriate material, and to provide clear reporting mechanisms for users. In the summer of 2024, the enforcement of these regulations led to the imprisonment of individuals for social media posts deemed to incite violence during protests, igniting national debates over the boundaries of online speech and state intervention. Critics argue that while the Act aims to create a safer online environment, it may inadvertently suppress legitimate expression due to its stringent measures. As in many other countries we look into in this report, the UK faces an ongoing challenge in reconciling the protection of its citizens with the preservation of fundamental freedoms in the digital realm.

Structure of the report

This report explores public attitudes towards social media, content moderation, and the delicate balance between freedom of speech and protection from harm. Organized around key questions concerning mostly the moderation of harmful speech (including hateful speech and mis- or disinformation), it seeks to illuminate how people across diverse cultural and political contexts perceive these complex issues. Each section assessed a specific aspect of the broader debate, offering comparative insights and highlighting the nuanced trade-offs that define public opinion.

1 Who Should Set the Rules? Public Attitudes on Responsibility for Content Moderation

This section asks who citizens believe should bear the primary responsibility for moderating content on social media. Should it be the platforms themselves, governments, or independent organizations?

2 Freedom or Protection? Navigating the Balance Between Speech and Harm

In this section, we explore how individuals balance the competing priorities of freedom of speech and the need for protection from harmful speech. We examine the cultural and political factors that shape these attitudes across different countries. The section also delves into how people define freedom of speech in their own terms, highlighting variations in interpretation and the boundaries they perceive as acceptable.

3 Toxic by Default? The Normalization of Hate and Harm in Online and Offline Spaces

Examining the prevalence and normalization of hate speech and toxicity, this section explores how exposure to harmful content shapes societal attitudes and behaviors both on and offline.

4 The big trade-off: Moderation, Misinformation, and the Desire for Safe Spaces

This section investigates how people view the trade-offs between effective content moderation and ensuring a digital space free from hate and misinformation, balanced with a desire for open discourse.

Methodology

This study has been commissioned by the Chair of Digital Governance at the Technical University of Munich (Prof. Yannis Theocharis) and the Department of Politics and International Relations at the University of Oxford (Prof. Spyros Kosmidis) to understand public attitudes towards platform content moderation and freedom of speech. It was conducted in collaboration with members of the Chair of Digital Governance and the Content Moderation Lab of the TUM Think Tank. The fieldwork took place between October 24 and November 26, 2024 and was coordinated by the German office of Bilendi & Respondi.

Samples were assembled using nationally representative quotas for age, gender, and education level. Specifically, we applied age quotas with five categories, gender quotas (male/female), and education level quotas based on the ISCED classification (three levels). These quotas ensured the representation of different demographic groups in each country. The total sample size is N=13,590. All graphs are based on the full sample. For an overview of the sample size for each country, along with population size and Internet penetration, please refer to the table below.

For South Africa, soft quotas were applied to age, gender, and education due to limitations in data availability. Similarly, in Brazil, age quotas for individuals 55 and older were difficult to implement, requiring the relaxation of quotas for these age groups in these markets. The target population consisted of individuals between 16 and 69 years of age, residing in the respective countries. The main sample achieved an incidence rate of 95–100% among the general population. To ensure data quality, two attention checks were incorporated. Respondents who failed either check were excluded from the survey.

The data were weighted to match demographic targets based on census and industry-accepted data for each country. This included adjustments for age, gender, and education to ensure that the sample accurately reflected the demographic structure of each population. In countries where Internet penetration is lower (e.g., South Africa), the data were interpreted as representative of the online population, rather than the national population, due to biases introduced by the mode of data collection.

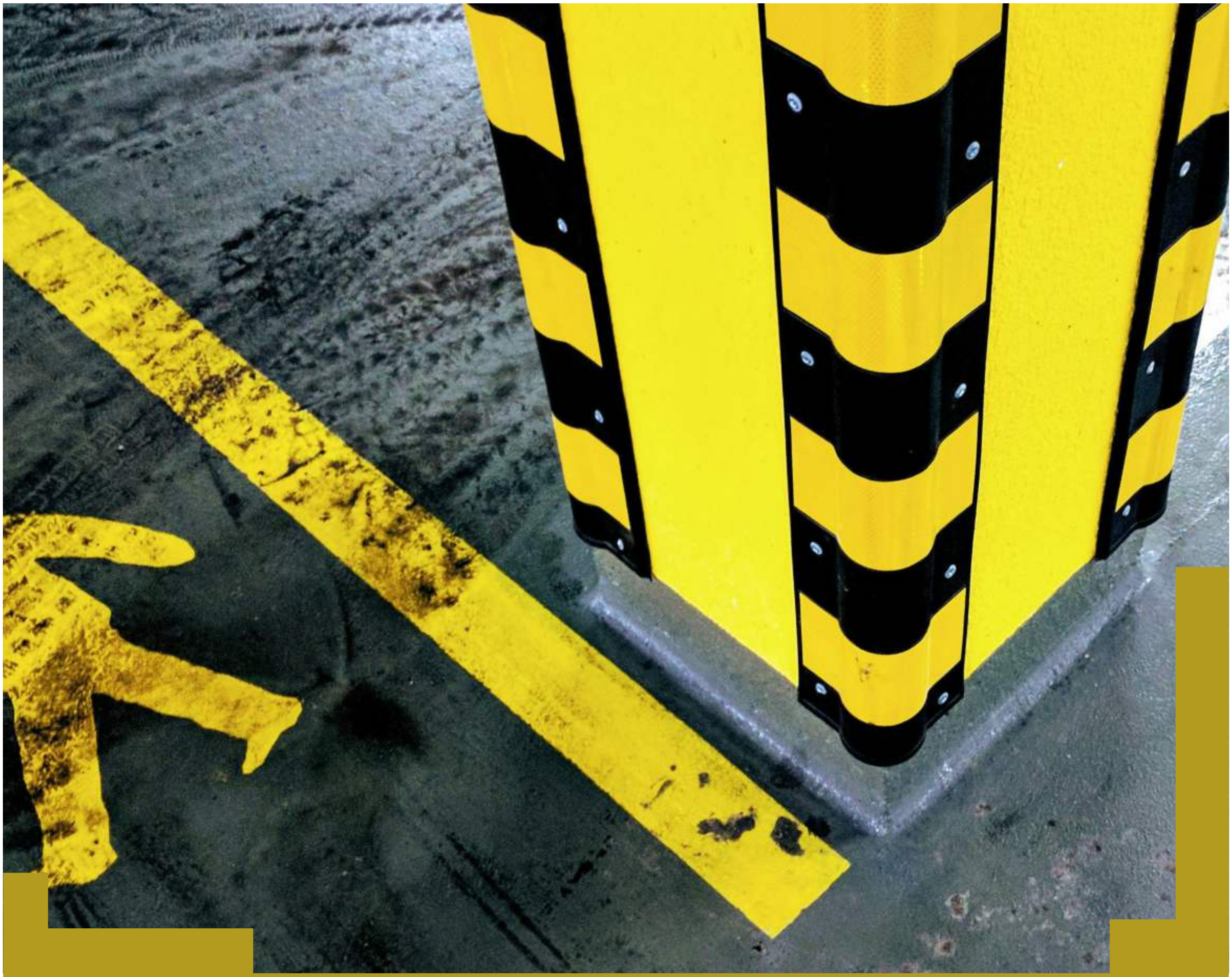
As with all online surveys, the results of this study are subject to limitations inherent in the data collection methodology. First, the study's reliance on online panels means that certain demographic groups may be underrepresented. Furthermore, the use of self-reported behavior introduces biases related to imperfect recall and social desirability, which should be considered when interpreting the findings. Although the sample sizes were relatively large, the ability to meaningfully analyze certain minority groups was limited.

Finally, for countries where quotas were less strictly enforced (e.g., South Africa and Brazil), caution is advised when comparing results across markets. In these instances, the differences between the online population and the national population may be more pronounced. To ensure transparency, data from India were excluded from the report and analysis due to identified concerns with data quality. We hope to expand the list of participating countries in future scientific publications and subsequent reports.

COUNTRY	SAMPLE SIZE	POPULATION	INTERNET PENETRATION
Australia	1328	26.5m	95%
Brazil	1374	211.1m	84%
France	1395	66.4m	87%
Germany	1373	84.5m	92%
Greece	1383	10.2m	85%
Slovakia	1389	5.5m	87%
South Africa	1360	63.2m	75%
Sweden	1309	10.6m	96%
United Kingdom	1349	68.7m	95%
United States	1330	343.5m	97%

Sample size, population, and intenet penetration per country

Source: Population size and Internet penetration using the latest available data for each country based on World Bank data..



1 Who Should Set the Rules?

Public Attitudes on Responsibility for Content Moderation

The responsibility for ensuring safe online spaces is a topic of growing concern as harmful content like hate speech and misinformation continues to spread. Citizens across democracies have differing views on whether platforms, governments, or individuals should be responsible for content moderation. These varying perspectives highlight the complexity of balancing freedom of expression with the need to combat online harm, offering valuable insights into how different societies approach the challenge of digital governance.

To understand how citizens across different democratic contexts view responsibility for content moderation, we designed two complementary survey questions that probe different aspects of platform governance and online safety.

The first question we posed to respondents across 10 countries directly addressed the fundamental issue of responsibility for maintaining online safety: "Who should be primarily responsible for maintaining a healthy and safe online environment?" Respondents were asked to choose between social media platforms, government regulation, individual citizens, or none of these options. This question helps us understand where the public places primary accountability for ensuring digital spaces remain conducive to healthy discourse.

Our second question approached the issue from a more specific angle, asking "Who is in the best position to combat harmful speech online?" Here, we expanded the range of options to include non-governmental organizations and civil society, while also specifying that individual citizen action could take the form of counterspeech. This question moves beyond abstract responsibility to assess who citizens believe can most effectively address harmful content in practice.

Together, these questions allow us to measure several crucial aspects of public opinion on content moderation. First, they reveal whether **citizens view content moderation as primarily a corporate responsibility, a government function, or a collective social obligation. Second, they help us understand if people make distinctions** between general platform safety and the specific challenge of combating harmful speech. Third, by including civil society organizations in the second question, we can assess whether citizens see a role for independent oversight beyond traditional institutional actors.

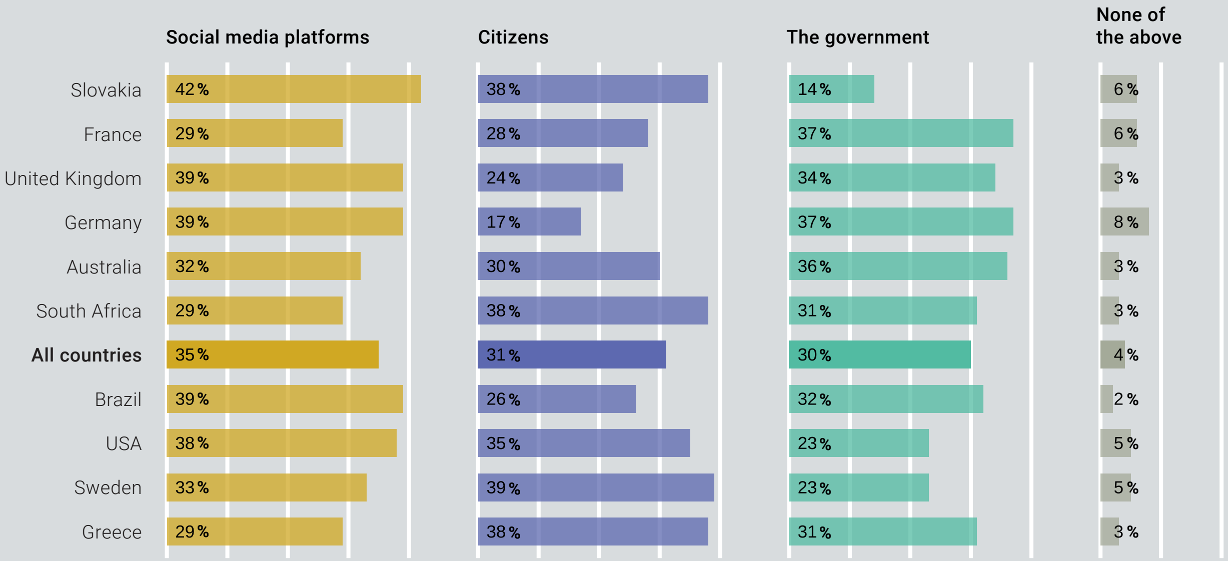


Figure 1.1 **“Who should be primarily responsible for maintaining a healthy and safe online environment?”**

Options (single-select): Social media platforms (e.g., Facebook, Twitter (X), etc.), The government through regulation or law enforcement, Individual citizens, None of the above.

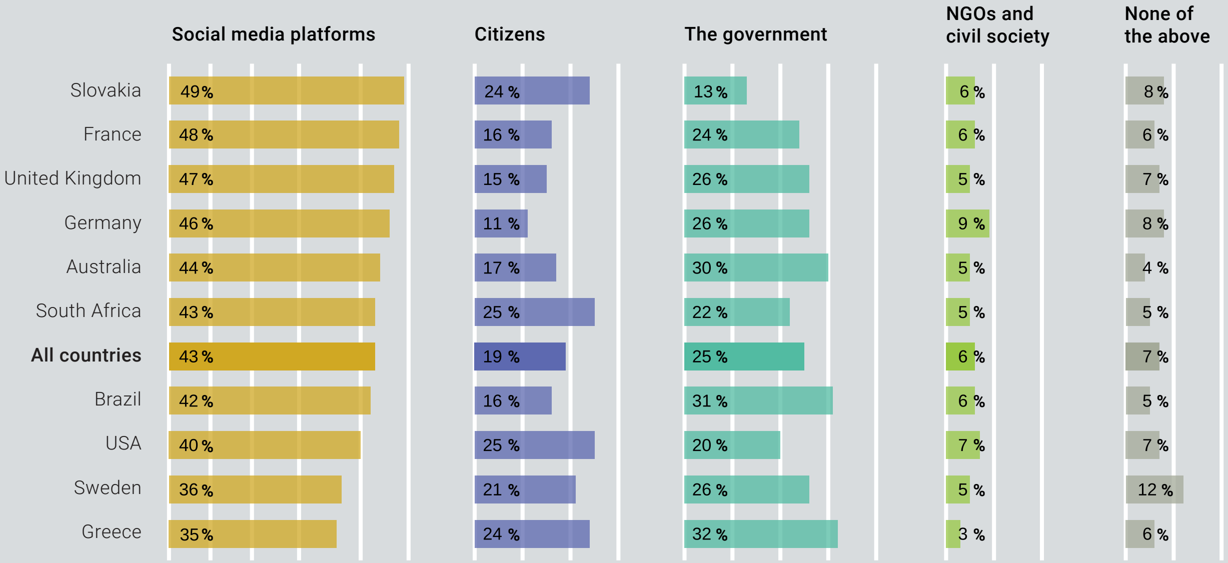


Figure 1.2 **“Who is in the best position to combat harmful speech online?”**

Options (single-select): Social media platforms (e.g., Facebook, Twitter (X), etc.), The government through regulation or law enforcement Individual citizens through counterspeech (e.g., responding to posts), Non-governmental organizations and civil society, None of the above.

This measurement approach is particularly valuable given the evolving nature of content moderation debates. As outlined in our introduction, platforms, governments, and civil society organizations are all actively engaged in shaping online governance and norms, but there has been limited systematic evidence about which of these actors citizens trust to perform this crucial function. Understanding these preferences across different national contexts can inform policy approaches that align with public expectations while highlighting potential gaps between institutional practices and citizen preferences.

The cross-national scope of our study allows us to examine how different regulatory traditions, political cultures, and experiences with online harm might influence public attitudes toward content moderation authority. For instance, we can explore whether citizens in countries with stronger state regulation of speech, such as Germany, differ in their views from those in countries with more laissez-faire approaches, like the US.

Nearly 35% of respondents believe social media platforms should bear primary responsibility for content moderation.

In Figure 1.1, our analysis reveals several striking patterns in how citizens across different countries view responsibility for maintaining a healthy and safe online environment. Looking at the aggregate picture across all countries surveyed, social media platforms emerge as the most frequently chosen responsible party, with nearly 35% of respondents believing they should bear primary responsibility. The top option is followed by individual citizens at 31% and government regulation at 30%. Only a small fraction (4%) rejected all proposed options.

However, this aggregate picture masks significant country-level variations. In Sweden, for example, there is a notably strong preference for citizen responsibility, with 39% of Swedes placing primary responsibility on individual users—the highest proportion across all surveyed nations. This citizen-centric view is also prominent in Slovakia (38%), South Africa (38%), and Greece (37%).

In contrast, several countries show a clear preference for government oversight. France leads this group with 37% favoring government responsibility, followed closely by Germany (37%) and Australia (36%). This preference for government involvement likely reflects these countries' stronger regulatory traditions in media governance.

Social media platforms are seen as the primary responsible party in several countries, particularly in Slovakia (42%), the UK (39%), and Brazil (39%). The US also shows a relatively strong preference for platform responsibility at 38%, despite its traditionally more laissez-faire approach to media regulation.

Perhaps most notably, Germany stands out for having the lowest proportion of respondents (17%) who believe individual citizens should bear primary responsibility, significantly below the cross-national average. This finding is particularly interesting given Germany's strong regulatory framework for online content moderation through laws like the Network Enforcement Act (NetzDG). (But we note that NetzDG has largely been repealed due to the adoption of the EU-wide DSA regulations, so regulation across the European Union will increasingly become standardized).

These results paint a complex picture of public attitudes toward online safety responsibility, suggesting that citizens across different democracies have varying expectations about who should take the lead in ensuring healthy online discourse.

When asked specifically about who is best positioned to combat harmful speech online, our data reveals a clear preference for platform-led solutions across most countries, though with notable variations in the degree of support and alternative preferences (Figure 1.2).

Social media platforms emerged as the preferred actor in combating harmful speech.

Social media platforms emerged as the preferred actor in combating harmful speech, with 43% of respondents across all surveyed countries viewing them as best positioned for this task. This preference is particularly pronounced in several European nations, with Slovakia (49%) and France (48%) showing the strongest support for platform-led approaches, followed closely by the UK (47%) and Germany (46%). Although the respondents in Greece were the least likely to indicate that social media companies were best-placed to counter online toxicity, the proportion of Greeks choosing this answer was only about 18 percentage points lower compared to Slovakia.

Government intervention received the second-highest level of support overall (25%), though with substantial cross-national variation. Relatively high rates of perceptions that the government is best-placed were observed in Greece, Brazil and Australia (but we note that even in Greece the plurality of respondents had more faith that social media companies could counteract harmful speech, if they chose to do so.)

Citizen-led approaches through counterspeech garnered significant but lower levels of support (19% across all countries). However, this average masks important regional variations. Several countries showed notably higher support for citizen involvement, with South Africa (25%), the US (25%), and Greece (24%) all recording levels of support above the international average. In contrast, Germany showed particularly low support for citizen-led approaches (11%).

Non-governmental organizations and civil society received consistently low levels of support across all surveyed nations, never exceeding 9% in any country. Germany showed the highest support for NGO involvement (9%), while Greece registered the lowest (3%). This relatively low support is noteworthy given the significant role that civil society organizations often play in monitoring and reporting harmful content. (The proportion of respondents selecting "none of the above" remained relatively low across most countries, though Sweden stood out with the highest percentage (12%) rejecting all proposed options.)

Overall, the patterns in Figure 1.2 represent an interesting contrast to the first question, suggesting that citizens may view the general maintenance of online safety differently from the specific challenge of addressing harmful speech. These findings suggest that while platforms are widely seen as best positioned to combat harmful speech, there is no universal consensus on the most effective approach, or even on who bears most responsibility for problematic content.

There is no consensus on the most effective approach or who bears responsibility.

2 Freedom or Protection?

Navigating the Balance Between Speech and Harm

The balance between free speech and harm prevention remains a central issue in the digital age. While the US emphasizes free expression, Europe focuses on accountability and harm prevention, as reflected in the Digital Services Act. Public views highlight how cultural and legal contexts shape opinions on where to draw the line between expression and safety, shedding light on the complexities of this ongoing challenge.

The public debate over content moderation and freedom of expression has intensified following Mark Zuckerberg's recent announcement to replace Meta's fact-checkers with community notes as part of a shift back to the platform's "free speech roots." With this, concerns have been raised around users losing even more trust in social media platforms, potential surges in false information and harmful speech, but also risks of offline violence fueled by online content (New York Times, 2025). This raises again the central issue of whether to prioritize the protection of freedom of expression or the prevention of harm. At the heart of this debate is whether removing content that some support, but which causes harm to others, infringes on the speaker's right to freedom of expression.

While the announcement has primarily focused on US values of safeguarding freedom of speech in alignment with the First Amendment, the issue carries global significance. In the US, nearly all forms of speech are constitutionally protected, whereas in Europe, laws explicitly prohibit discriminatory speech against minorities and incitement to violence. The legal traditions of the two regions reflect distinct approaches to balancing freedom of expression with protecting individuals from harm (Kohl, 2022). This contrast is evident in recent European Union legislation, such as the Digital Services Act, which emphasizes accountability and harm prevention in digital spaces (European Commission, 2024). Both traditions regard their models as optimal for preserving free speech as a cornerstone of democracy and fostering a marketplace of ideas.

Examining the European approach to content moderation also reveals notable differences among countries. Germany stands out as a pioneer, having introduced the Network Enforcement Act, the first hate speech law of its kind, later replaced by the broader Digital Services Act. This legislation served as a blueprint adopted by other nations, such as France, which implemented even stricter measures for moderating online content (Heldt, 2019). In contrast, Slovakia presents a different case: a country grappling with political extremism and violence, it initially resisted robust content moderation but has since shifted its regulatory approach following the consequences of inaction (Švec et al. 2024).

In this section, we first focus on how the public views freedom of speech as a core societal value. We then explore how these attitudes might shift when respondents are confronted with the trade-off between prioritizing freedom of speech and prioritizing protection from harm. Finally, we examine respondents' perspectives on the various nuances of freedom of speech. Given the subjectivity and cultural differences that shape attitudes towards free speech, we asked the respondents to indicate their level of agreement with a range of statements that capture the facets of freedom of expression, both online and offline. This analysis will offer a more nuanced understanding of the complex interplay between speech protection and harm prevention, shedding light on how societies navigate this critical issue in the digital age.

Freedom of Speech as a core value

Does the public view freedom of speech as a core societal value that should be upheld, even when it means offending others? The survey results reveal intriguing patterns in how strongly respondents value freedom of speech (Figure 2.1) and their tolerance for offensive speech (Figure 2.2). Countries such as Greece and Germany show strong agreement with freedom of speech as a core value, but strikingly, their willingness to tolerate offensive speech remains more moderate.

The US, as it was perhaps to be expected, stands out as an exception. While it aligns closely with the average across all countries (74%) in prioritizing freedom of speech, it ranks highest in agreeing that people should be free to express themselves, even if it hurts, offends, shocks or disturbs others (53%). This reflects the deep-rooted cultural and legal traditions in the US that strongly favor protecting individual speech rights, even in the face of controversy.

In contrast, respondents from Australia and the UK exhibit both lower support for freedom of speech as a core value and less tolerance for offensive speech, reflecting their regulatory frameworks and stances, which prioritize protection from harm, as shown in Figure (FOS). Meanwhile, respondents from South Africa and Brazil express strong support for freedom of speech as a core value, but their views become more divided when it comes to whether free speech should be protected even when it offends others (Figure 2.2). In Brazil, only 30% of respondents agree that free speech should be protected in such cases, the lowest among the countries surveyed. On the other hand, South Africa, which generally leans toward prioritizing protection from harm, shows a higher agreement at 45%, indicating that they see free speech as a core value and a significant portion of respondents support free speech, even when it leads to offense. This stands against their general emphasis toward protection from harm, especially in light of the second statement.

Overall, the results highlight no clear-cut link between freedom of speech values and behavior and reveal a complex, often contradictory, relationship between support for free speech as a core value and the importance of not offending others with it. This suggests that public attitudes toward speech rights are shaped by more than just abstract values, including cultural, legal, and political contexts.

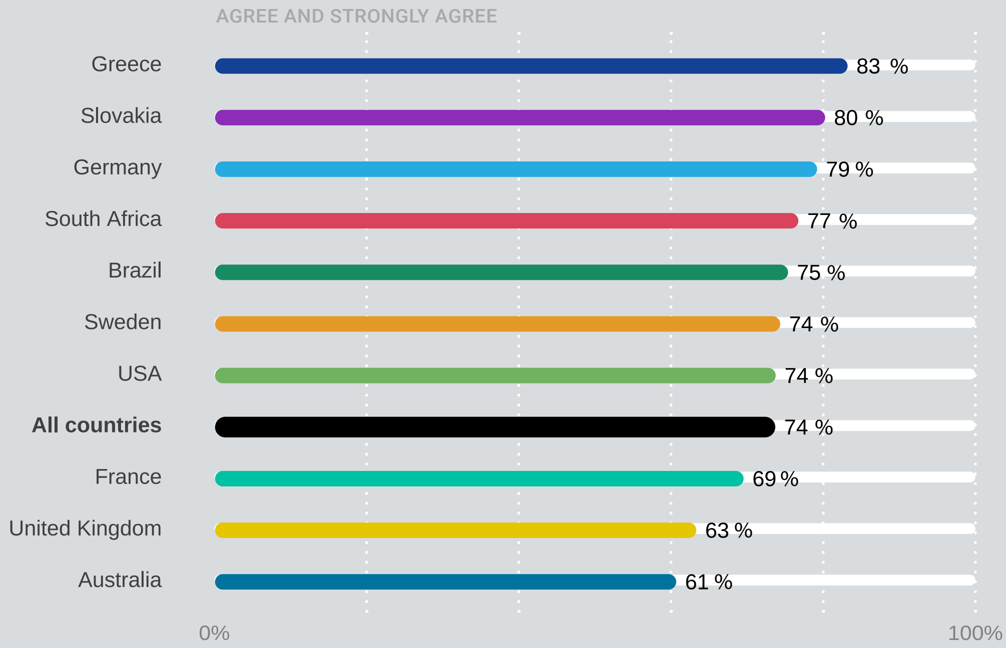


Figure 2.1 **Prioritize free speech as a core value**

“Please indicate the extent to which you agree with the following statements: We should prioritize free speech as one of the most important values in our society.”

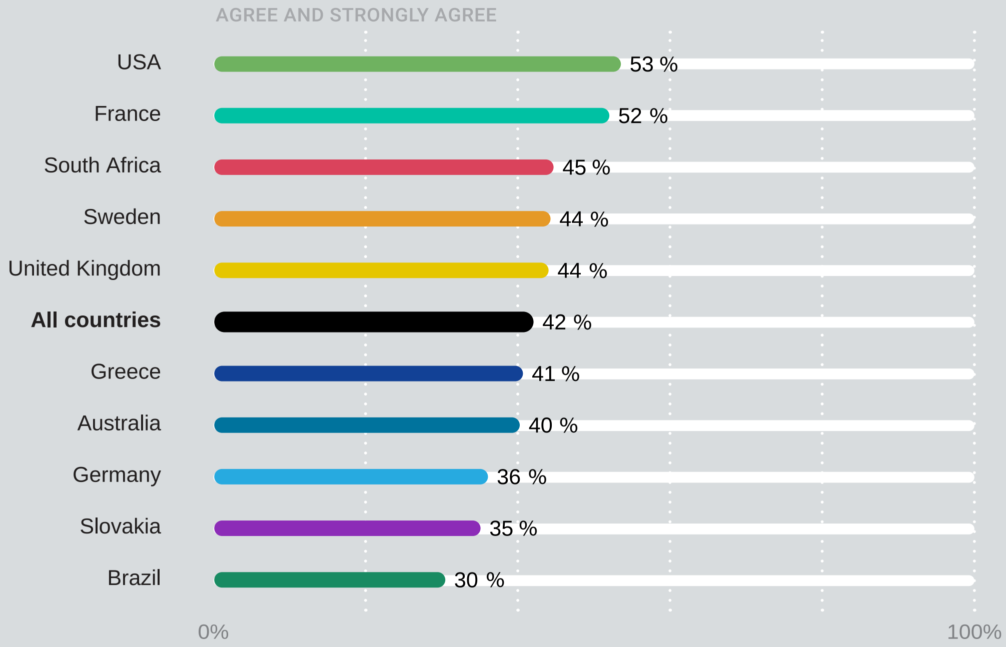


Figure 2.2 **Free speech is important even if it offends others**

“Please indicate the extent to which you agree with the following statements: We should be free to express ourselves, even if hurts, offends, shocks, or disturbs others.”

Shown are the percentages of those who either ‘agree’ or ‘strongly agree’

How Important is Freedom of Speech Compared to Protection from Harm?

While freedom of speech is widely recognized as a core societal value, its application is often more complex in practice. People frequently face a trade-off between upholding free speech and protecting individuals from harm, which challenges the notion of freedom of speech as an absolute value. To better understand public perspectives on this issue, we posed a critical question: "In general, how important is freedom of speech relative to the harm it might cause?" This question, deliberately designed without references to social media platforms or specific content examples, serves as a measure of public sentiment across countries. By analyzing the responses, we gain insights into how national traditions and policies regarding harmful content and misinformation shape public attitudes, revealing this trade-off valued by users in each country.

Figure 2.3 illustrates the findings from our analysis of user preferences across ten democracies regarding the balance between protecting freedom of speech and preventing harm. The results reveal stark differences in national tendencies, shaped by varying legal frameworks and cultural traditions.

At one end of the spectrum, Sweden stands out as the country with the strongest preference for protecting freedom of speech, with an average score of 37.4 on a scale ranging from 0 ("Protecting Freedom of Speech") to 100 ("Protection from Harm"). Greece follows with an average score of 41.8, closely aligned with the US, which also averages at 42.4. Interestingly, despite the strong emphasis on freedom of speech enshrined in the US Constitution through the First Amendment, US respondents in our sample show a more moderate tendency. The US ranks third, showing a tendency to balance the protection of speech with the prevention of harm, rather than prioritizing absolute free speech.

Germany presents a particularly noteworthy case. Despite its legal tradition of stringent hate speech laws, the average score for German respondents is 43.1, placing them just behind the US and suggesting a slight lean toward protecting freedom of speech. However, the distribution of responses in Germany indicates a more normalized and balanced spread across the scale, consistent with the country's legal framework and the European tradition of balancing freedom of expression with safeguards against harmful speech.

On the opposite end of the spectrum, South Africa emerges with the highest average score of 56.3, reflecting a pronounced preference for safeguarding individuals from harm. Brazil follows closely with an average score of 51.1, while France's score of 49.5 also places it at the center. The UK is almost identical in distribution to France with an average of 49.4 and similar variation across responses. Following similar patterns are Australian respondents' attitudes with an average of 49. In South Africa and Brazil, the broader distributions suggest a diversity of opinions within the population, yet both countries consistently lean toward a prioritization of harm prevention—likely related to both countries' legacies of social inequality and racial discrimination. In contrast, France's more concentrated distribution reveals a preference for striking a middle ground between the protection of free speech and the need for harm prevention. This balanced stance resonates with broader European traditions, where regulating harmful speech is viewed as essential for maintaining human dignity.

While both Brazil and South Africa exhibit higher average scores, indicating a stronger inclination toward harm prevention, it is particularly striking that South African respondents emphasize this protection with consistency, underscoring the country's distinct attitudes on the debate complex issues.

Slovakia also comes up as a very interesting case. The country shows a relatively balanced distribution of responses, reflecting a preference for striking a balance between protecting freedom of speech and preventing harm. This balanced stance comes in contrast to Slovakia's political landscape, which has historically included strong advocates against content moderation. However, recent events, including instances of political extremism and violence, seem to have shifted public opinion toward a more moderate approach, suggesting a growing recognition of the need for regulation to prevent harm.

Overall, our survey results highlight significant differences across the ten countries in balancing freedom of speech with the prevention from harm. These differences reflect how national legal

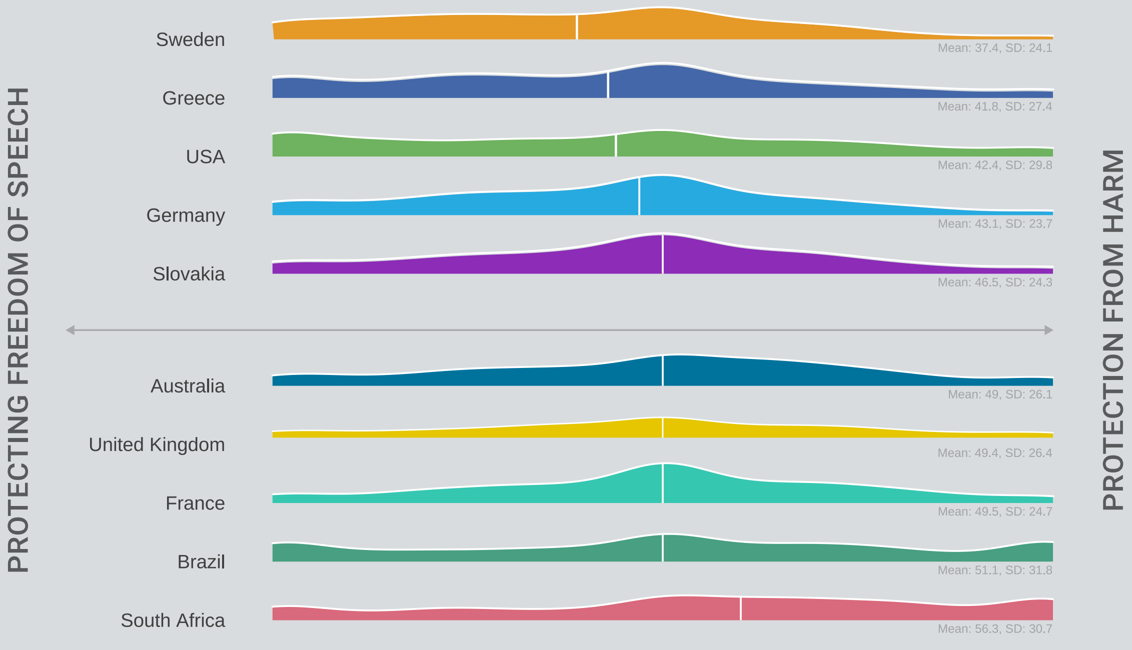


Figure 2.3 **Importance freedom of speech vs. harm it might cause**

"In general, how important is freedom of speech relative to the harm it might cause?" Options (single-select): Scale 0= 'Protecting Freedom of Speech'—100= 'Protection from harm'. The vertical lines show the median, and the statistics show the mean and the standard deviation (SD) for each country distribution.

traditions, political experiences, and cultural norms may shape public preferences on this issue. While countries like Sweden, Greece, the US, and Germany lean more toward protecting freedom of speech, others, such as South Africa, Brazil, and France, display a stronger tendency toward harm prevention. Notably, several countries—including Slovakia, France, and even the US—show a substantial share of respondents who favor striking a balance between the two extremes.

While some countries lean more toward protecting freedom of speech, others display a stronger tendency toward harm prevention.

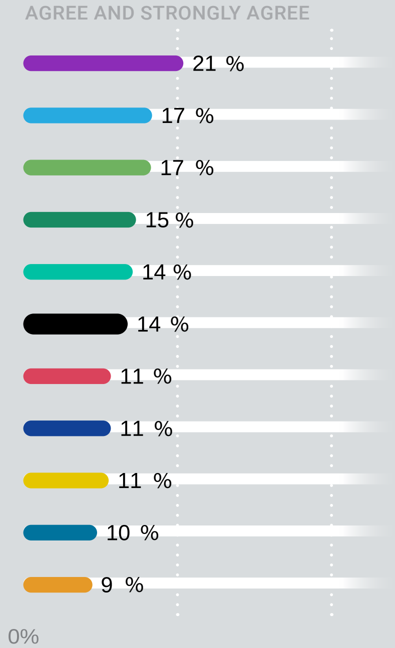
These findings are particularly interesting in light of ongoing debates about content moderation, often led by influential figures like Mark Zuckerberg and Elon Musk. Both have argued that freedom of speech should take precedence to uphold democratic values and strengthen human rights. However, our data suggests that most respondents across these ten democracies prefer a more balanced approach. This runs counter to the argument that prioritizing unrestricted free speech will inherently protect democratic values. Instead, there is growing recognition by the public that failing to moderate harmful content can further marginalize minority voices and undermine democratic principles by allowing harmful speech to flourish unchecked.

The limits of freedom of expression

Next, we presented respondents with statements addressing different dimensions of freedom of speech, including opinions that threatening posts should be permitted online to enable counter-speech, the belief that regulation is the only means to limit hate speech and that speech inciting violence should be banned.

Our findings reveal that when it comes to different dimensions of free speech, such social media platforms should allow posts threatening others with violence to stay online, so that users can respond and counteract them with counter-speech (Figure 2.4), public support is generally low. Only 14% of respondents across all countries agree that threatening posts should be allowed to stay online to give users a chance to respond with counter-speech. The highest support is seen in Slovakia (21%), followed by Germany (17%), and the US (17%), while Sweden (9%) and Australia (10%) show the least support, indicating that most people prefer removing threatening content altogether rather than leaving it up for counter-speech. Notably, respondents from Sweden scored lowest on the scale, indicating the highest support for protecting freedom of speech (Figure 2.3). This clearly shows the importance of looking into the facets of freedom of speech and considering the contextual nature of it. It seems that protection of free speech is very important to some but seen as less important to protect when crossing this specific boundary.

In contrast, we can see more variation in views on whether regulation is necessary to limit hate speech (Figure 2.5). While 45% of the overall sample agree that only regulation can effectively limit harmful speech, countries like France (69%), South Africa (55%), and Brazil (53%) show the strongest support for this statement. This aligns with the respondents' broader attitudes toward the balance between protecting freedom of speech and preventing harm. Notably, France's score stands out, diverging significantly from the overall average and reflecting a strong preference among French respondents for regulating harmful content. On the opposite end, Sweden (32%)



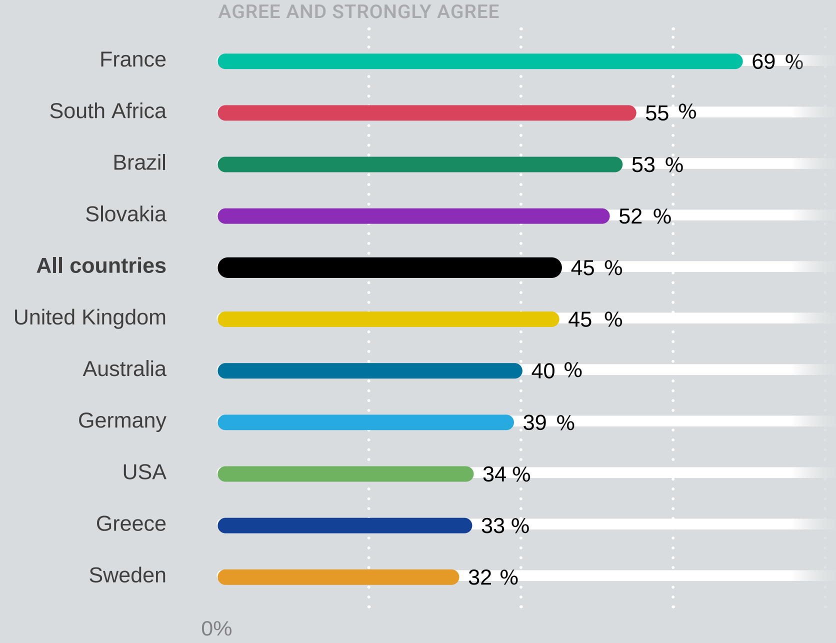


Figure 2.4 **Allow threatening posts online for counter-speech**

"Please indicate the extent to which you agree with the following statements: Social media platforms should allow posts threatening others with violence to stay online, so that users can respond and counteract them with counter-speech."

Figure 2.5 **Only regulation can limit hate speech**

"Please indicate the extent to which you agree with the following statements: Only regulation can limit hateful speech."

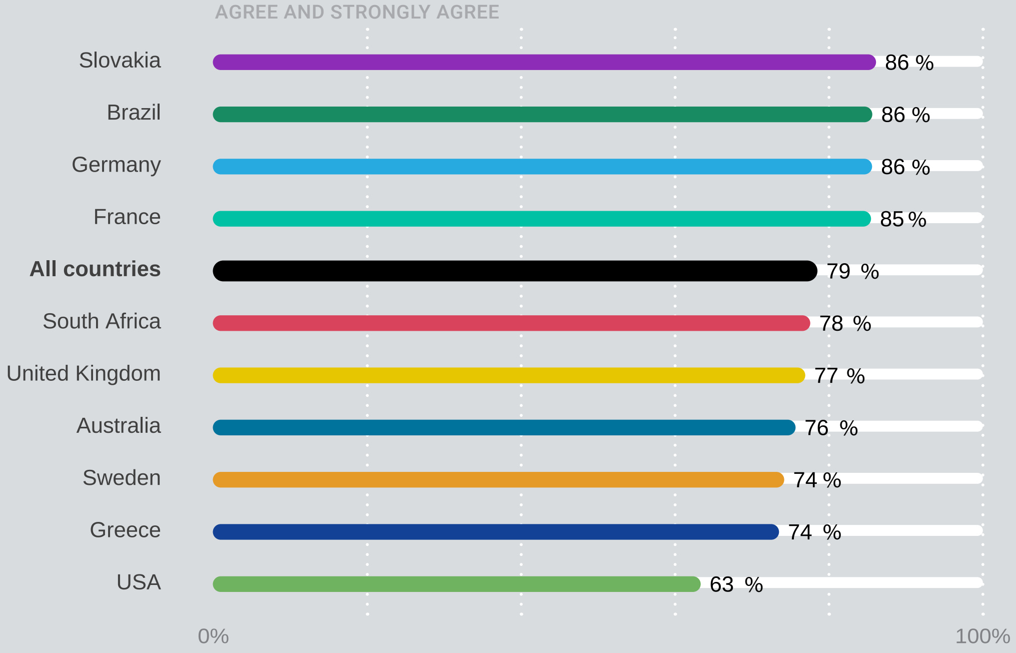


Figure 2.6 **Ban speech inciting violence**

"Please indicate the extent to which you agree with the following statements: Speech inciting violence should be banned."

Shown are the percentages of those who either 'agree' or 'strongly agree'

and Greece (33%) show less support, indicating a more skeptical view on government intervention in online speech.

Despite these differing views on regulation, there is broad consensus across countries that speech inciting violence should be banned (Figure 2.6). On average, 79% of respondents support banning such speech, with Slovakia, Brazil, and Germany leading at 86%, followed closely by France at 85%. Additionally, respondents from South Africa (78%), the UK (77%), and Australia (76%) show strong agreement. The US stands out as a notable outlier, with 63% of respondents supporting a ban on speech that incites violence—still a clear majority, but significantly lower than in other countries. This reflects the US's unique legal protections for free speech, which tend to be more robust, even in extreme cases, compared to other democracies. Nonetheless, the broader pattern across all countries is clear: a majority agrees that speech inciting violence should be banned, indicating a general preference for protection over the support of unrestricted free speech.

A majority agrees that speech inciting violence should be banned.

Overall, our results show that while countries vary in their approaches to regulating harmful speech, there is broad agreement on the need to remove violent content. However, the US stands apart, showing more resistance to restrictions on harmful speech. The findings suggest that the ongoing debate around counter-speech and regulation is shaped by national traditions—some countries prioritize stricter regulations to prevent harm, while others, like the US, focus on safeguarding speech rights, even if it means allowing potentially harmful content to remain online. Still, there is broader support for limiting free speech by the public to protect individuals, even though the legislation in some countries acts in opposition to these sentiments.

What about Free Speech Offline?

While much of the focus on free speech revolves around online spaces, we also explored attitudes toward free expression in offline contexts. The findings show mixed views, highlighting the complex nature of free speech both online and offline. To assess respondents' offline attitudes, we presented them with two statements: "Comedians should be allowed to say what they want without any restrictions." and "Libraries should not remove books with content that goes against our society's core values."

When asked whether comedians should have the freedom to speak without restrictions, 45% of respondents across all countries agreed (Figure 2.7). However, this still leaves the majority supporting some form of restriction, highlighting a general belief that comedy should have boundaries, particularly when it comes to sensitive or potentially harmful content. Slovakia stands out with the highest support for unrestricted speech by comedians at 54%, followed closely by the US, Greece, and the UK (all at 48%). On the other hand, Brazil shows the least support at just 28%, reflecting a more cautious stance on the limits of comedic freedom.

In the context of libraries, 54% of respondents believe that books opposing core values should not be removed (Figure 2.8). Support for keeping controversial books is strongest in Sweden (62%) and

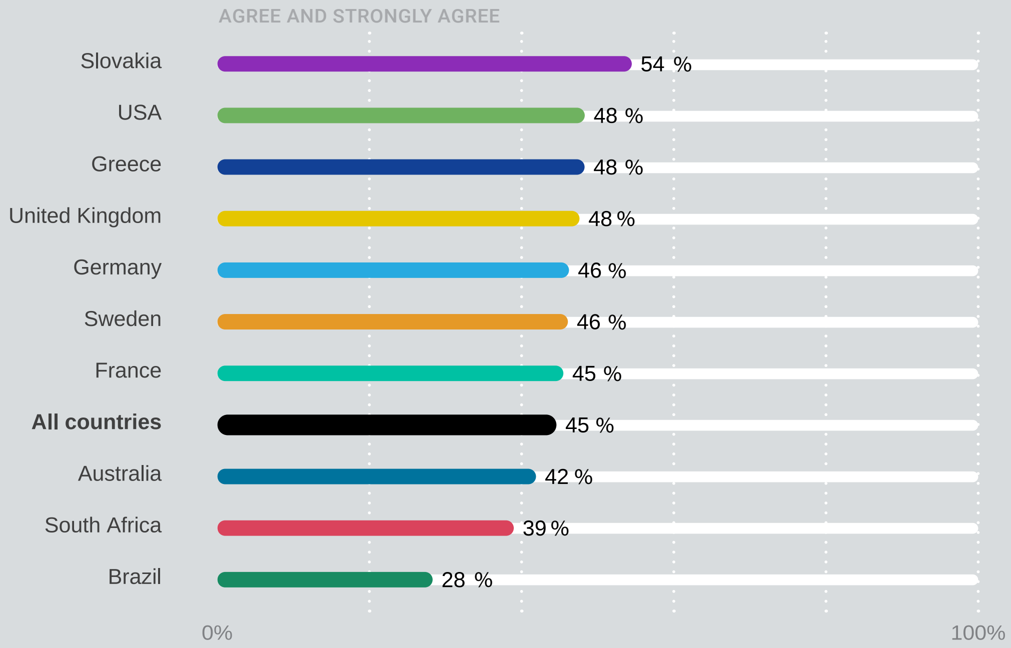


Figure 2.7 **Comedians should speak freely without restrictions**

“Please indicate the extent to which you agree with the following statements: Comedians should be allowed to say what they want without any restrictions.”

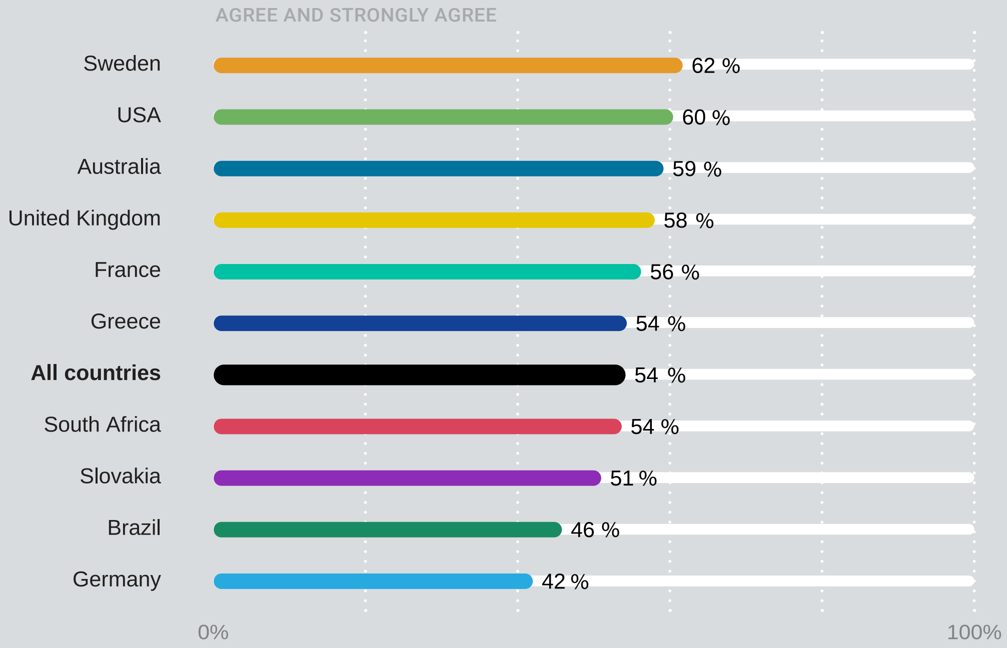


Figure 2.8 **Libraries should not remove books that oppose core values**

“Please indicate the extent to which you agree with the following statements: Libraries should not remove books with content that goes against our society’s core values.”

Shown are the percentages of those who either ‘agree’ or ‘strongly agree’ the US (60%). This reflects a strong belief in preserving access to diverse ideas and reflecting on the general sentiment in these countries to support free speech. It also provides interesting context for the fact that book bans in US schools and libraries surged in 2023 (New York Times, 2024). However, Germany stands out, with only 42% agreeing that such books should stay in libraries. Thus, in Germany, where there is a strong legal tradition against hate speech, people are more inclined to support removing books that could undermine societal values.

Taken together, these numbers suggest that while people generally value free speech offline, they see a need for more restrictions in certain contexts, particularly when it comes to comedians. At the same time, most respondents agree that libraries should remain spaces where diverse viewpoints, even controversial ones, are preserved.

Misinformation and toxicity

Finally, we presented respondents with a statement suggesting that people should be able to share information on social media that is considered false by the government, and that Internet users should be able to post offensive content about certain groups if they want to criticize them.

The results reveal considerable variation in attitudes toward misinformation and toxic online behavior. When asked whether people should be allowed to share false information even if the government disapproves, 44% of respondents across all countries agreed (Figure 2.9). Majorities in some of the countries support this stance with Greece (59%), Sweden (56%), and Slovakia (54%), scoring highest, indicating stronger protections for freedom of expression even in cases of misinformation. In contrast, countries like the UK (31%) and Brazil (32%) exhibit much lower agreement, indicating a more cautious stance on tolerating misinformation. Interestingly, US respondents score below the international average, with only 40% in agreement, challenging the popular belief that US values align strongly with the First Amendment's emphasis on free speech. Thus, when considering this particular trade-off, it appears that US respondents and those from countries below the overall average are more reserved when confronted with false information.

However, attitudes toward hate speech and toxic content differ strikingly (Figure 2.10). When asked whether users should be allowed to post offensive content to criticize groups, only 17% of our respondents across all countries agreed. The US (29%) stands out as the country with the highest support for this stance, followed by France (22%) and Sweden (22%), indicating a stronger belief in the right to offensive speech, yet, still only a smaller fraction of the overall sample. In most other countries, support remains very low, with Germany (15%) and South Africa (11%) showing minimal agreement.

These findings suggest that tolerance for misinformation and offensive speech varies across contexts. There also seems to be a particular trade-off between preferences towards free speech when it comes to misinformation and hate speech. While some countries lean more toward protecting the right to share false or controversial content, the broader consensus is that hate speech and toxic behavior should not be tolerated online.

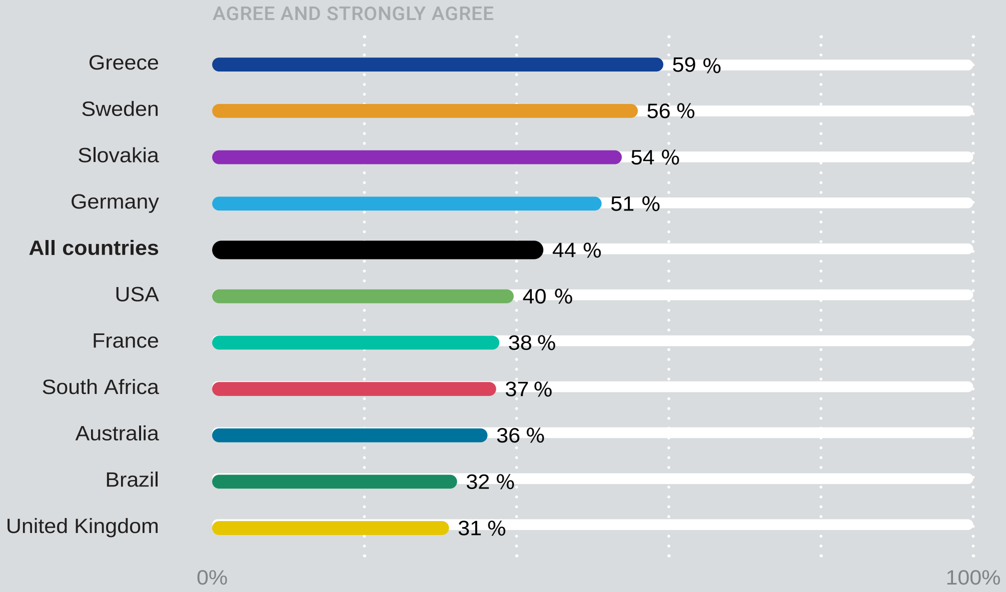


Figure 2.9 **People can share false information, even if the government disapproves**

"Please indicate the extent to which you agree with the following statements: People should be able to share information on social media that is considered false by the government."

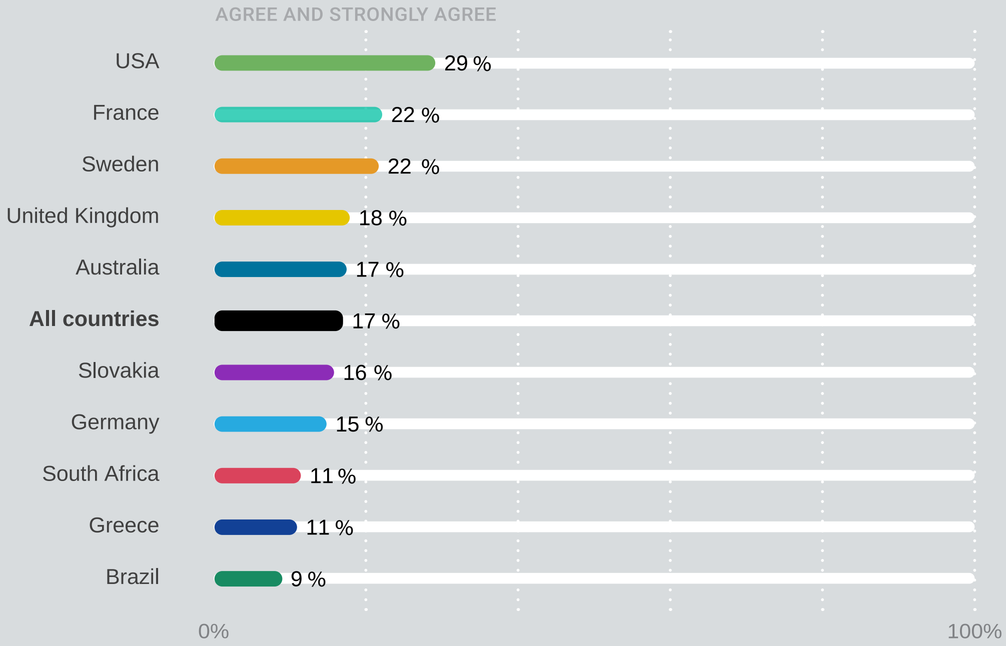


Figure 2.10 **Users should post offensive content to criticize groups**

"Please indicate the extent to which you agree with the following statements: Internet users should be able to post offensive content about certain groups if they want to criticize them."

Shown are the percentages of those who either 'agree' or 'strongly agree'

One of the key takeaways from the findings is the critical importance of employing a clear conceptualization and a multi-dimensional measurement strategy when examining public attitudes toward freedom of expression. Such an approach enables one to move beyond the broad, generalized ideas that citizens may associate with the concept of free speech. Often, belief in free speech functions as an abstract principle closely tied to national or cultural identity. As such, it may not be something people have fully contemplated in terms of specific, tangible scenarios. By presenting respondents with concrete dimensions of free speech—with real-world situations that may not have been immediately obvious to them as part of their understanding of the concept—we can better capture their actual perspectives. This approach helps bring nuance to a discourse often framed in binary terms: "freedom of speech versus restrictions of speech". It acknowledges the complex and sometimes contradictory relationship between support for free speech as a core value and the recognition of the importance of not offending others in its exercise. This approach helps to illuminate the boundaries respondents place on free speech, providing a deeper understanding of their values and priorities. These insights might otherwise remain hidden when relying solely on abstract or generalized questions, particularly given how culturally loaded such questions can be.

3 Toxic by Default?

The Normalization of Hate and Harm in Online and Offline Spaces

Toxic speech online is increasingly normalized, with exposure to hate speech impacting individuals and society. The prevalence of harmful content, particularly toward vulnerable groups, has led to a sense of resignation, with many users feeling that addressing hate is futile. This highlights the need for more effective solutions to tackle online abuse.

The proliferation of toxic speech online has sparked intense debate among scholars, policymakers, and the public about whether hate is becoming "normalized" in both online and offline spaces. Normalization, in this context, refers to the process through which behaviors once considered unacceptable are increasingly perceived as routine or unavoidable. Academic studies provide evidence that exposure to hate speech has both individual psychological and societal effects. Bilewicz and Soral (2020), for example, have shown that repeated exposure to derogatory language about immigrants and minority groups leads to political radicalization, deteriorates intergroup relations, and erodes empathy. Research on toxic speech in video games by Beres and colleagues (2021) also shows that while toxicity is a pervasive problem that harms players' well-being and enjoyment, many normalize toxicity as an inextricable and acceptable element of the gaming experience. Hate speech affects human functioning beyond intergroup relations. In a functional Magnetic Resonance Imaging (fMRI) study, Pluta and colleagues (2023) showed that exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain.

The normalization of hateful content poses a challenging problem for societies by undermining anti-discrimination norms, fostering contempt for outgroups, and desensitizing individuals to the offensive nature of such language and the pain it inflicts on its victims. If widespread, this process can create a dangerous feedback loop: societies become increasingly tolerant of derogatory language while growing more hostile toward the targets of such hate. These concerns have intensified as recent media reports and academic research highlight the pervasive presence of harmful content on social media platforms (Hate Aid, 2021). Studies, both by academics and non-governmental organizations like Amnesty International and Hate Aid, show that this content disproportionately impacts specific groups, including women, LGBTQ+ individuals, and ethnic minorities—groups that stand to benefit most from the incredible opportunities for free expression that social media offers. Compounding this issue is a complementary process: the lack of reporting of hateful content. Surveys indicate that few people report hate when they encounter it (Das Nettz et al., 2024). This

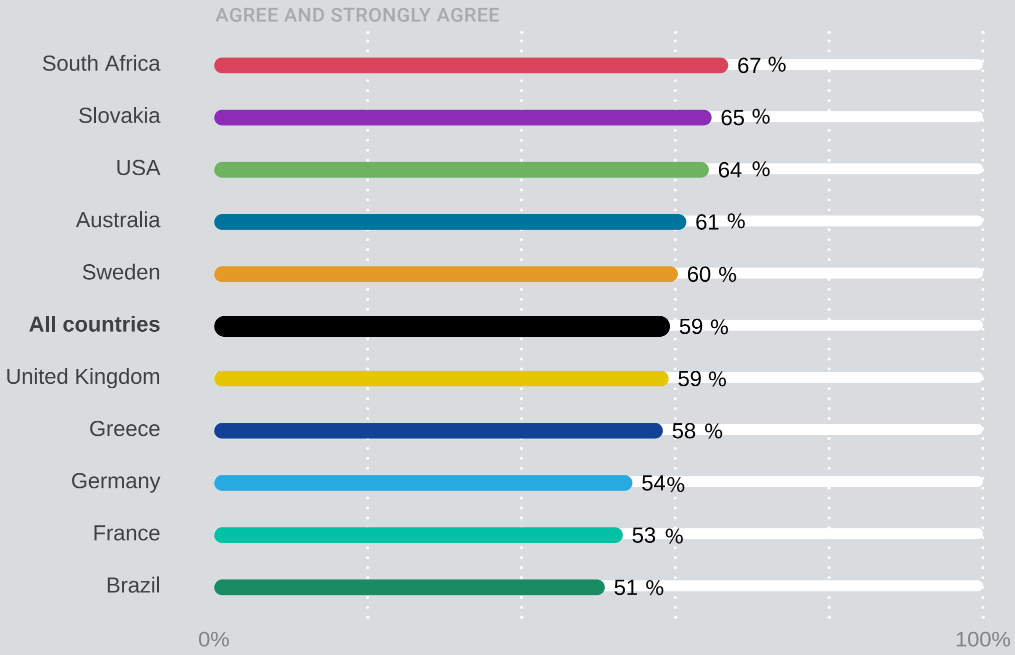


Figure 3.1 **Toxic language on social media is unavoidable**

“Please indicate the extent to which you agree with the following statements: Exposure to toxic language (either incivility, intolerance, or hate) across social media platforms is unavoidable”.

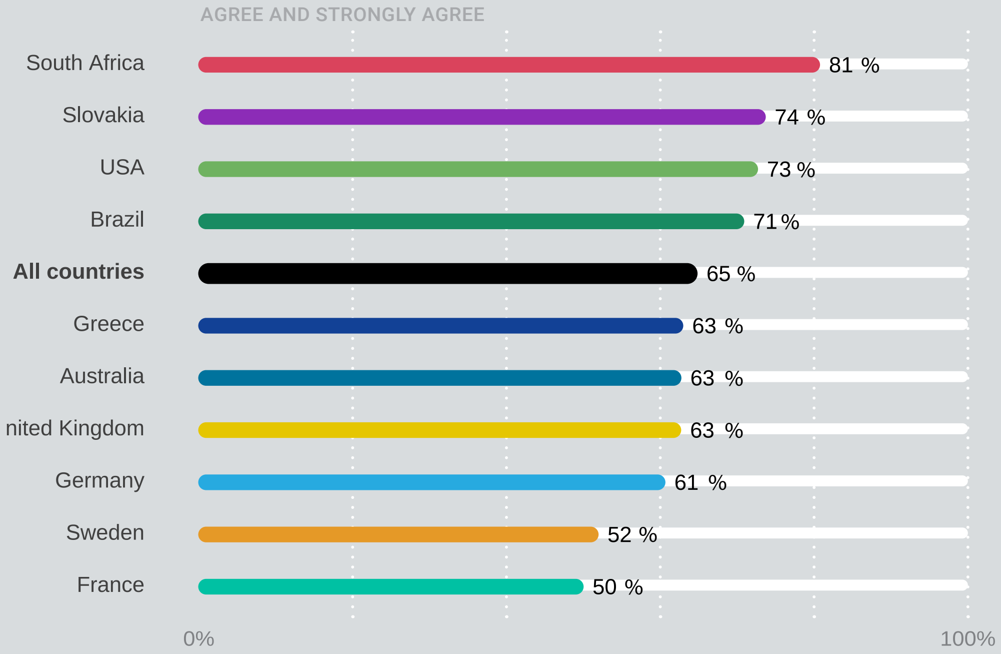


Figure 3.2 **Sharing views invites aggressive replies**

“Please indicate the extent to which you agree with the following statements: If you share your views on social media, you must be ready for aggressive replies from those who disagree.”

Shown are the percentages of those who either ‘agree’ or ‘strongly agree’.

underreporting raises a critical question—one that has been difficult to answer due to a lack of empirical evidence: Have we become so accustomed to hate and harm that it feels futile—or even unnecessary—to push back?

In this section we present important findings from public perceptions of the normalization of hate, shedding light on the ways people experience and respond to toxic environments, both online and offline.

We begin with a number of statements aimed at understanding people's perceptions of how common manifestations of online toxicity are and gauging the extent to which they believe these behaviors have become normalized. The statements "Hate speech, intolerance, and incivility online are so common it's hard to imagine that social media platforms will improve" and "Exposure to toxic language (either incivility, intolerance, or hate) across social media platforms is unavoidable" allows us to explore whether individuals feel that the pervasiveness of hate speech, intolerance, and incivility on social media platforms has reached a point where meaningful improvement seems unlikely.

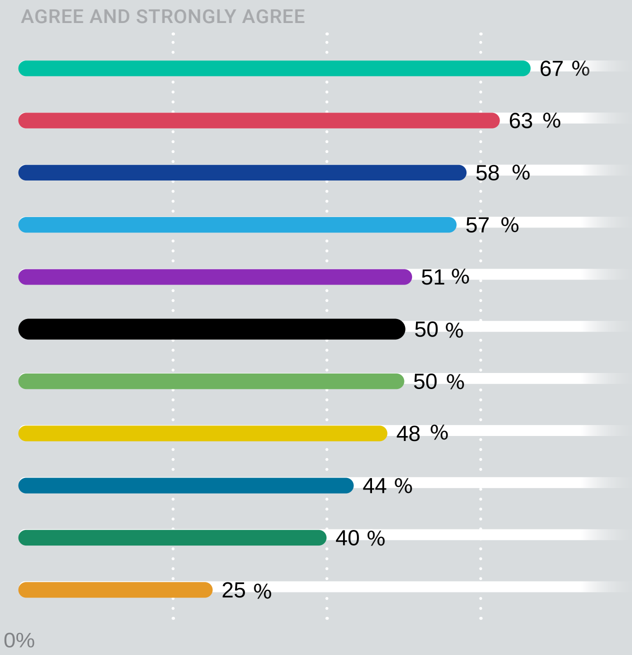
As can be seen in Figure 3.1, while there is some variation across countries in terms of how people feel that this type of speech is common in online platforms (from 51% in Brazil and 53% in France, to 65% in Slovakia and 67% in South Africa), the fact that majorities across countries agree that these are all common and unavoidable types of speech on social media platforms hints at a broader sense of resignation that many users feel, suggesting that, despite widespread calls for change and platforms' past promises to combat hate, the scale and persistence of harmful content make it hard to envision platforms successfully addressing the issue.

The sheer scale and endurance of harmful content makes it hard to imagine platforms effectively combatting it.

The widespread acceptance of aggressive exchanges on online platforms is reflected again in our dataset, where 65% of respondents agree with the statement, "If you share your views on social media, you must be ready for aggressive replies from those who disagree" (Figure 3.2). Notably, this expectation is particularly pronounced in countries like South Africa, where more than 80% of respondents anticipate such behavior, compared to around 50% in France. In the US, where recent evidence from the Pew Research Center indicates that public opinion largely believes political debates have become less respectful (84%) (Pew Research Center, 2023), almost 73% of respondents expect their views to be met with aggressive content.

This growing expectation of hostility online seems to reflect broader societal trends. Our survey also found that in many countries respondents believe rudeness and disrespect are prevalent in day-to-day interactions outside of the Internet (Figure 3.3). In France, almost 67% of respondents agreed with this statement, while 63% in South Africa and 50% in the US shared similar views. While this does not seem to be the case everywhere (only 25% of Swedes agreed that people's behavior offline is rude), on average people across countries agree that "It is very common in large gatherings to see uncivil behavior when two or more people disagree on an issue" (Figure 3.4). The overall trend, thus, suggests that the normalization of toxic behavior

The normalization of toxic online behavior is part of a larger pattern of incivility.



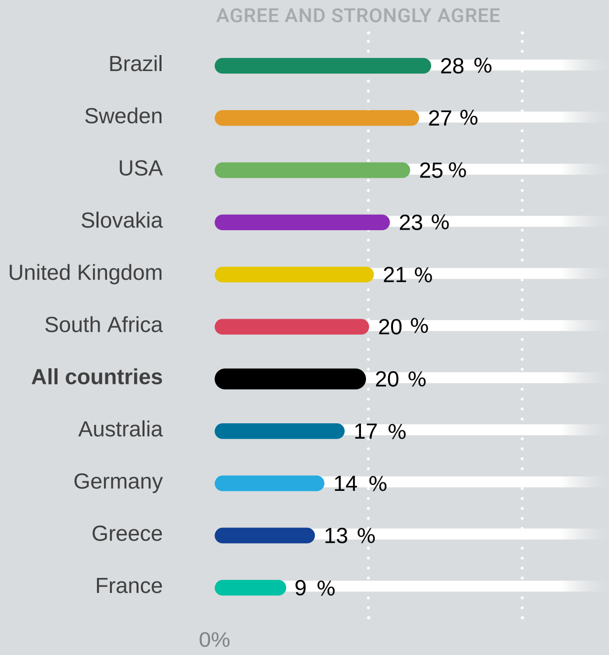


Figure 3.3 **People are often rude offline**

"Please indicate the extent to which you agree with the following statements: In our day-to-day interactions outside the Internet, people tend to be rude and disrespectful"

Figure 3.4 **Uncivil behavious is common in disagreements**

"Please indicate the extent to which you agree with the following statements: It is very common in large gatherings to see uncivil behavior when two or more people disagree on an issue".

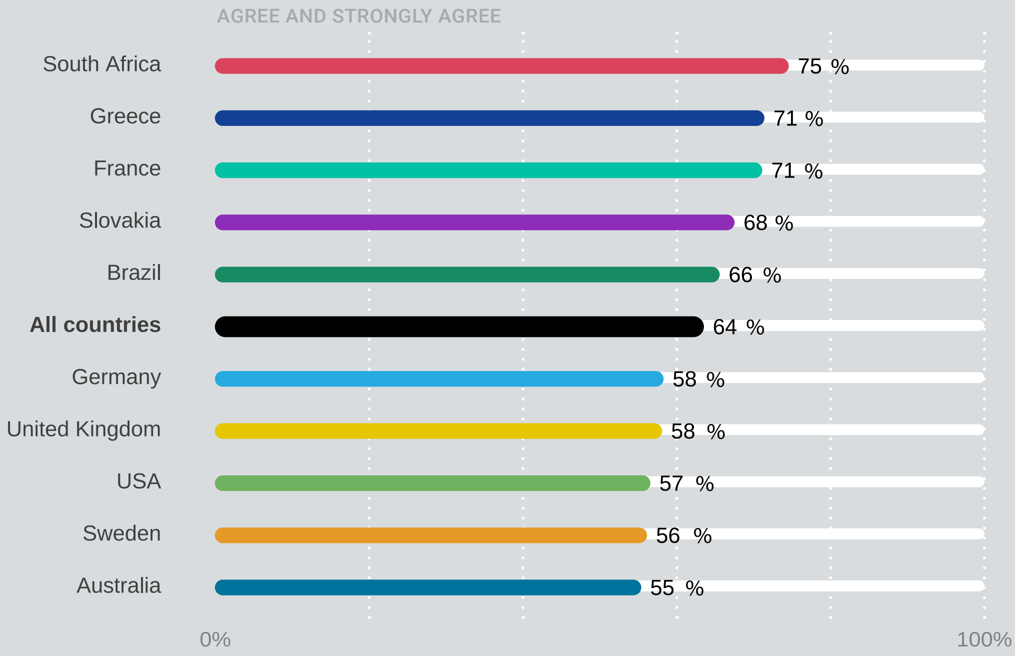


Figure 3.5 **Rudeness is sometimes needed on social media**

"Sometimes you need to be rude on social media to get your point across"

Shown are the percentages of those who either 'agree' or 'strongly agree'

online is not an isolated phenomenon but part of a larger pattern of incivility across both digital and physical spaces.

Our comparison suggests that across various societies, aggressive interactions are increasingly perceived as an inevitable part of social engagement, both online and offline, rather than something that can be curbed or prevented. Yet, while perceptions of others' online and offline behavior tend to lean toward the view that they are uncivil, we find clear evidence that such behavior is not necessary for people to express their views on social media.

Across all the countries we studied, only about 20% of survey participants agreed with the statement, "Sometimes you need to be rude on social media to get your point across," with around 28% in Brazil and only 9% in France (Figure 3.5). In short, perceptions of normalization of hostile behavior do not equate to acceptance. Despite the widespread expectation of aggression, particularly on social media, the majority of respondents across our study still reject the notion that rudeness is necessary to effectively communicate. While toxic behavior may be pervasive, it is not in any way a universally embraced or desired mode of interaction.

Rudeness isn't necessary to communicate effectively, according to respondents.

User experiences

While the perception of online toxicity as unavoidable is widespread, it is also crucial to understand how these dynamics manifest in users' lived experiences. Online toxicity has garnered significant attention in the mainstream media in recent years due to multiple reports highlighting its sharp rise (see, for example, the Pew Research Center's 2021 report "The State of Online Harassment", which tracks online abuse through surveys since 2017). For instance, X made headlines shortly after billionaire Elon Musk acquired the platform and laid off many staff members, including the head of its content moderation team. This led to a rapid increase in identity-based abuse and especially anti-semitic content (which was already on the rise), which was widely reported by journalists (Dwoskin et al. 2023). While some celebrated the change, others questioned whether the surge in hateful content made it ethical to continue using the platform (Williams, 2024). Much of the media coverage concerning increasing levels of toxicity on social media revolves around people's personal experiences, and especially of those belonging to specific groups. Across the countries we examine, beyond agreeing that hate speech, incivility, and aggression are common, many respondents reported facing direct harm in online spaces. In this section, we delve into personal accounts of online victimization, highlighting the types of harassment, abuse, and discrimination that users told us they encounter. From offensive name-calling to more severe incidents such as physical threats and doxing, our survey captures the breadth of harmful behaviors affecting individuals across different demographics. Furthermore, we examine how these experiences are often tied to specific aspects of identity, including gender, political views, religion, race or ethnicity, and sexual orientation.

I'VE BEEN CALLED OFFENSIVE NAMES

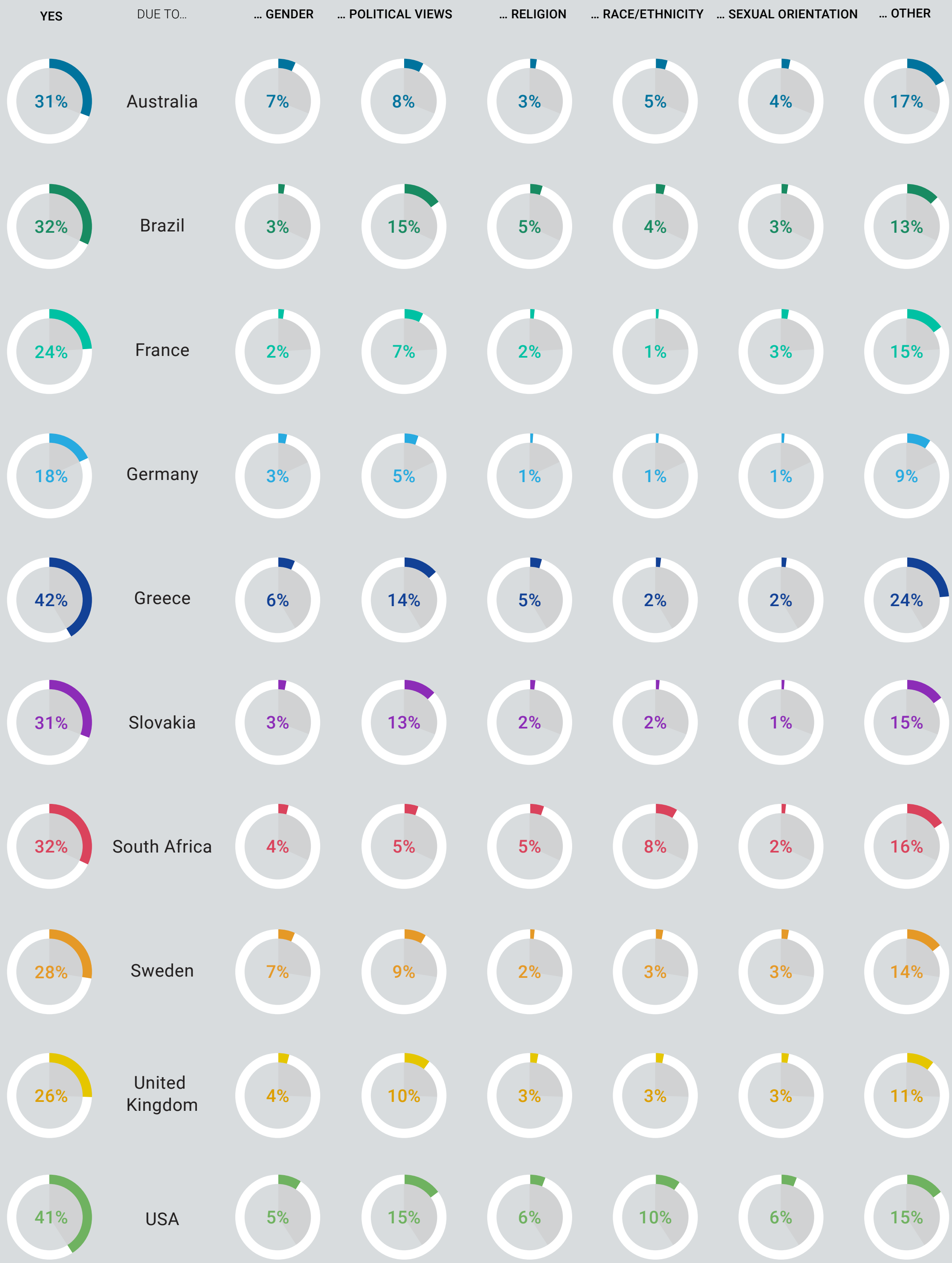


Figure 3.6 **Being called offensive names online and identified factors**

Percentage of respondents who reported being called offensive names ("I have been called offensive names") and who identified gender, political views, religion, race/ethnicity, sexual orientation, or other reasons as factors in being called offensive names. ("Which, if any, of the following have happened to you, personally, online?")

Respondents could select more than one reason where applicable; all percentages

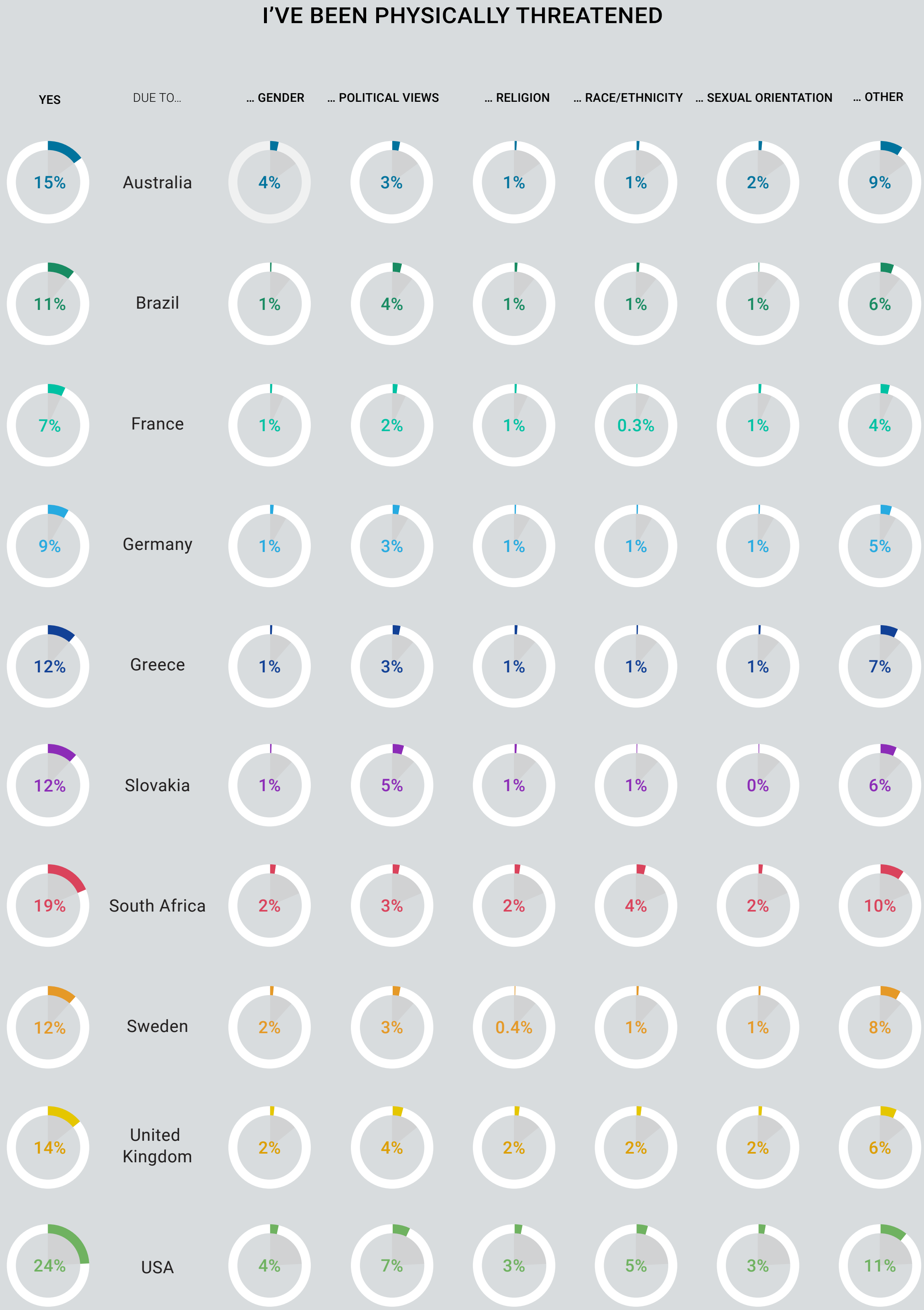


Figure 3.7 **Being physically threatened online and identified factors**

Percentage of respondents who reported being physically threatened online ("I have been physically threatened online") and who identified gender, political views, religion, race/ethnicity, sexual orientation, or other reasons as factors in having received physical threats online.

Respondents could select more than one reason where applicable; all percentages

BEING CALLED OFFENSIVE NAMES

One of the most common forms of online harassment reported in our survey was being called offensive names. Across the 10 countries studied, a significant proportion of respondents shared that they had personally experienced this form of verbal aggression in digital spaces, underscoring the pervasive nature of incivility online (Figure 3.6). Specifically, more than 40% of Greek and American social media users reported being called offensive names online, compared to less than 25% in Germany and France. Following the Pew Research Center's strategy in their "The State of Online Harassment" reports, for each type of online abuse we inquired about, we also asked participants to indicate the reason behind the attack, providing options such as gender, political views, religion, race or ethnicity, sexual orientation as well as "other reasons".

Expressing political views is a more common reason for online abuse than gender identity, sexual orientation, or race or ethnicity.

While gender identity, sexual orientation, and race or ethnicity are often cited in the media as key reasons for online abuse, our data suggests that being called offensive names is more commonly linked to the expression of political views. While data on this topic outside the US is almost non-existent, this finding aligns with evidence reported by the Pew Research Center (2021), which found that of the roughly four-in-ten Americans who had reported experiencing online harassment, half cited politics as the reason they believed they were targeted. This highlights the significant role political polarization plays in shaping online harassment dynamics, not just in the US, but beyond. Figure 3.6 to Figure 3.8 present a detailed breakdown of the types of attacks people have experienced online and the reported reasons for being called offensive names, physically threatened, and discriminated in each country, providing insight into the key factors driving online harassment..

BEING PHYSICALLY THREATENED ONLINE

Online threats of violence are among the most serious forms of online abuse. In many of the countries examined in this report, they are considered criminal offenses and can be prosecuted as such. Given this context, it is not surprising that fewer people report experiencing this type of abuse compared to more common forms, such as offensive name-calling. However, in certain countries, the percentages are considerable (Figure 3.7). For instance, almost 25% of respondents in the US reported receiving physical threats online, followed by respondents in South Africa and Australia. In contrast, only just over 5% of respondents in Germany and France had experienced this type of abuse. Interestingly, while political views continue to be a relevant factor in receiving physical threats online—along with race and ethnicity, particularly in South Africa, the US, and somewhat less so in the UK—most people indicated that they received physical threats for reasons outside the factors examined in our questionnaire.

DISCRIMINATION

Discrimination is another deeply harmful form of online abuse that disproportionately affects marginalized groups. It can manifest in various ways, including targeted harassment, exclusionary behavior, and verbal attacks based on a person's identity—such as their gender, race, ethnicity, sexual orientation, or religion. Our dataset reveals significant variations across countries in the prevalence

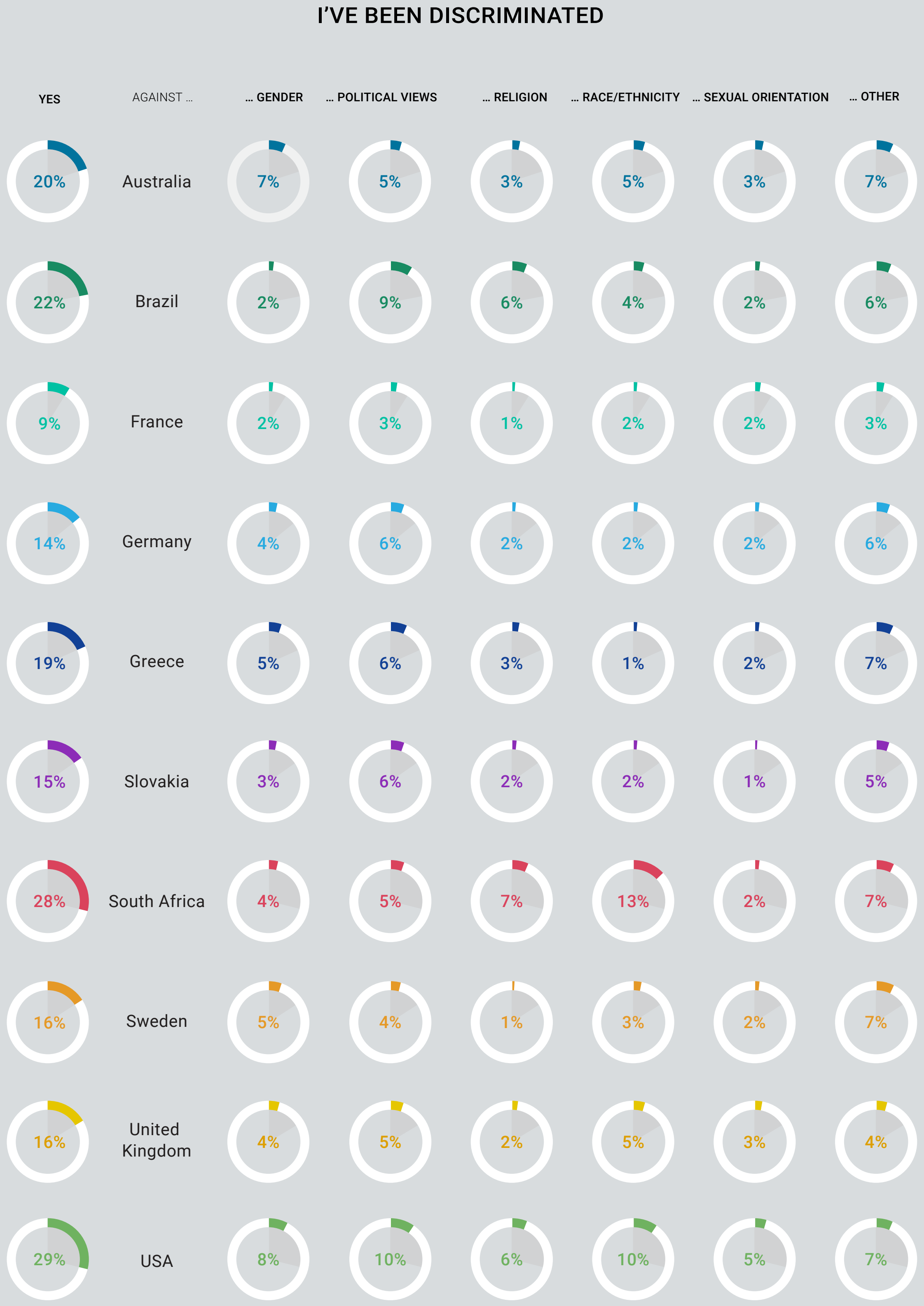


Figure 3.8 **Being discriminated online and identified factors**

Percentage of respondents who reported experiencing online discrimination in a given country who identified gender, political views, religion, race/ethnicity, sexual orientation, or other reasons as factors in being discriminated against online.

Respondents could select more than one reason where applicable; all percentages

of this type of abuse, with Americans—followed by South Africans—reporting experiences of online discrimination about three times more frequently than the French and twice as often as the Germans. In the US, the primary reasons for discrimination are related to race or ethnicity, followed by political views and gender, with the US also reporting the highest percentage of this type of discrimination (Figure 3.8). In South Africa, whose national Human Rights Commission (2017) noted in 2017 that "racism, racial bias and racial discrimination expressed on social media platforms in South Africa is routine and pervasive", the primary cause of online discrimination is also race or ethnicity. This trend highlights that longstanding societal divides not only remain deeply entrenched, but that social media "provides a fertile breeding ground through which it [discrimination] manifests, and is proliferated, allowing for real-time widespread harm and further entrenching hatred". Discrimination based on political views is the leading reason for this type of abuse in Brazil, Greece, Slovakia, the UK, France, and Germany, while gender is the primary reason in Australia and Sweden.

SEXUAL HARASSMENT

Sexual harassment is another distressing form of online abuse, involving unwelcome sexual advances, remarks, or behavior. Recent estimates show that more than 300 million children fall victim to online sexual abuse each year (Siddique, 2024). Platforms have faced severe penalties for failing to adequately address this issue, with Twitter being fined approximately $600,000 in Australia in 2024 for not meeting basic online safety expectations (Taylor, 2023). This form of abuse can have profound consequences on the victim's mental and emotional well-being, often leading to anxiety, depression, and a pervasive sense of powerlessness. Approximately 20% of Americans and Greeks have been sexually harassed online, followed by Brazilians, Swedish and Australians, all of whom stand just under 15% (Figure 3.9). Among our respondents, gender is the clear and dominant factor for being sexually harassed.

Gender is dominant factor in sexual harassment

DOXING

Of all the forms of online abuse explored in this report, doxing is the one that people reported experiencing the least. Doxing, which involves the release of an individual's private, personal information—such as their home address, phone number, or workplace— is considered a criminal offense in many of the countries we explore in this report can have devastating consequences, both online and offline. While it remains a relatively less common form of abuse in comparison to other types we explored in this report, its effects can be far-reaching, leading to real-world threats, harassment, online mobbing and serious privacy violations which may leave victims feeling exposed, vulnerable, and unsafe. A little over one in ten Americans have experienced this type of abuse online, followed by South Africans (11%) (Figure 3.10).

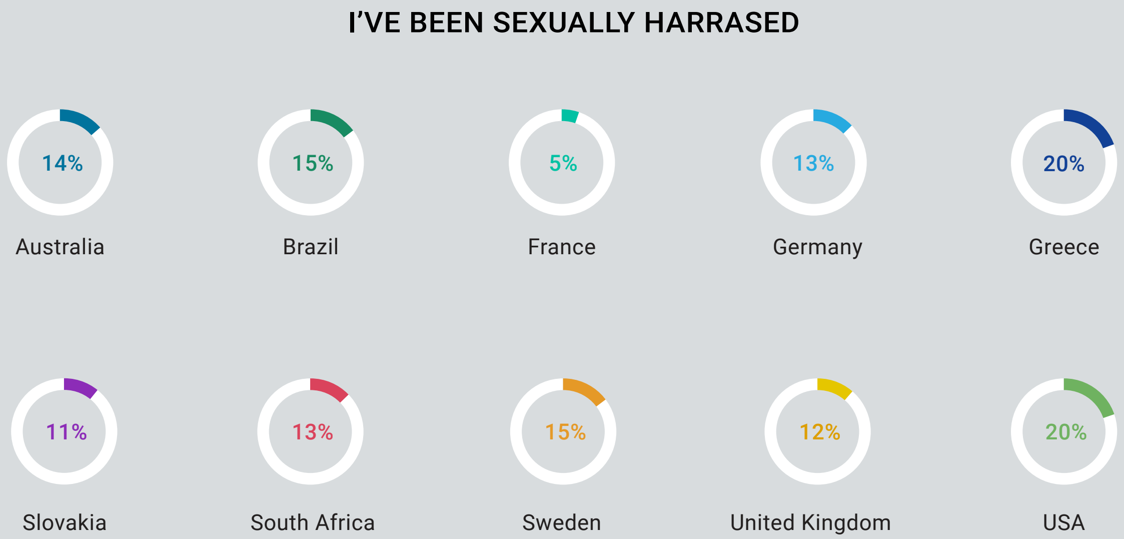


Figure 3.9 **Being sexually harrased online**

Percentage of respondents who reported being sexually harassed ("I have been sexually harassed").

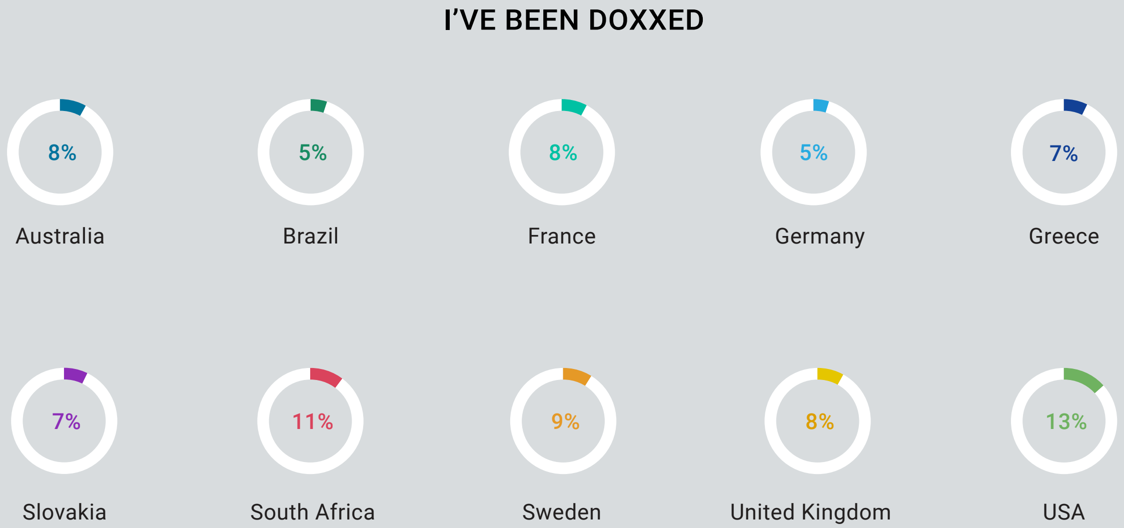


Figure 3.10 **Being doxxed online**

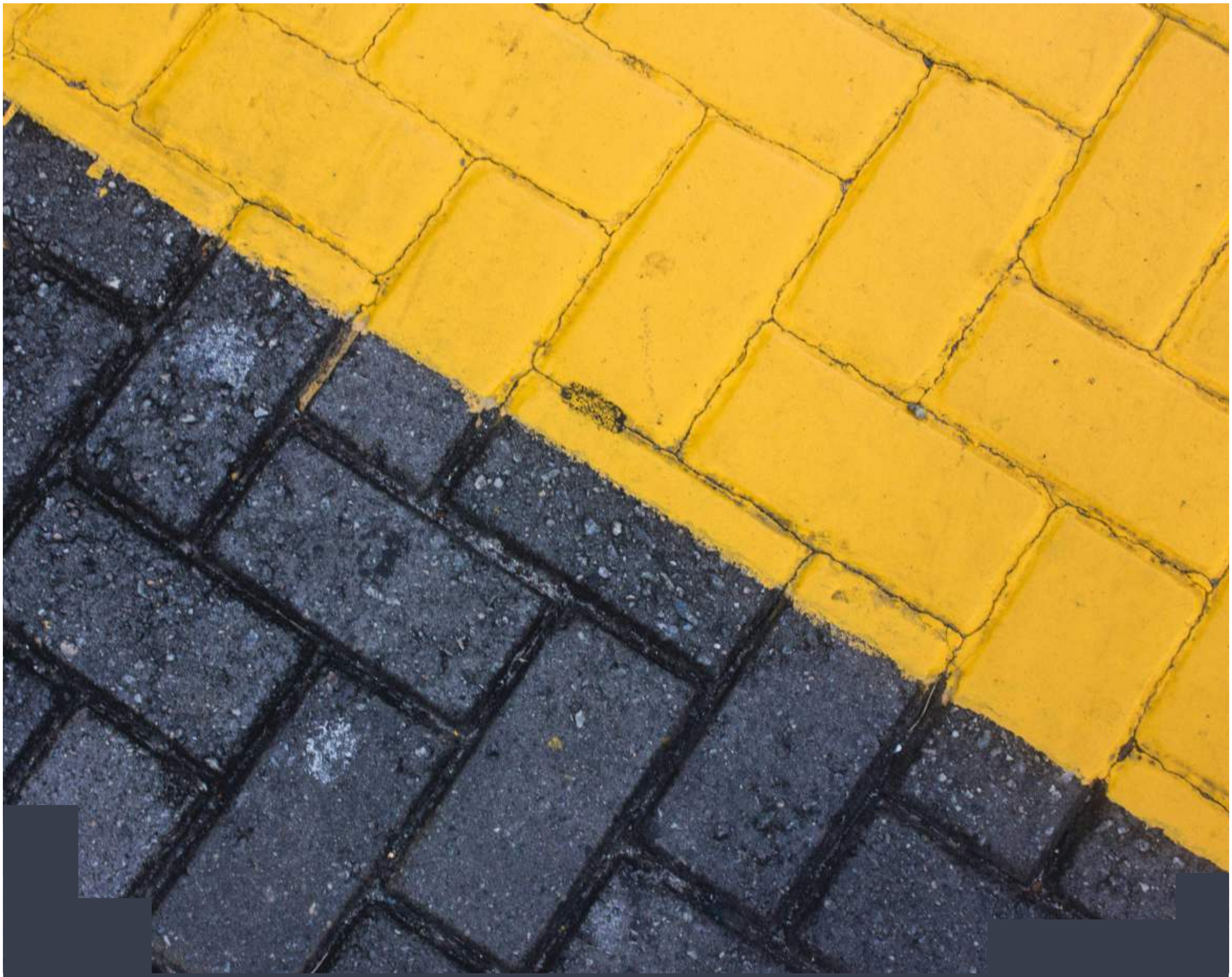
Percentage of respondents who reported being doxed ("I have been doxed (my private/personal material was published by someone on the Internet)").

Shown are the percentages of those who either 'agree' or 'strongly agree'

Summing up

In conclusion, the findings from both the perceptions and lived experiences of online users underscore the widespread normalization of toxicity across digital platforms. Our survey results reveal that a significant portion of respondents across the 10 countries studied perceive online hate, incivility, and discrimination as common and unavoidable, suggesting a deep sense of resignation about the capacity of social media platforms to address these issues effectively. This perception is not without basis, as the lived experiences of users reflect the prevalence of online harassment, with many reporting instances of verbal abuse, physical threats and discrimination. These harmful behaviors often stem from sensitive identity factors such as political views, race, gender, and sexual orientation, further highlighting the role of polarization in shaping online interactions. While some forms of abuse, such as being called offensive names, are more common, discrimination based on political views appears to be a significant contributor to harassment across multiple countries, particularly in the US and beyond. The impact of online abuse is far-reaching, affecting individuals' mental well-being and contributing to a culture of fear and disengagement.

The lived experiences of online users underscore the widespread normalization of toxicity across digital platforms.



4 The big trade-off

Moderation, Misinformation, and the Desire for Safe Spaces

The trade-off between no content moderation and a platform free from hate speech and misinformation shows that most respondents prefer moderation, particularly to reduce misinformation. While the US remains an outlier, globally, people value protection from harmful content over unrestricted posting.

While social media have been instrumental in amplifying important causes such as the #MeToo movement, the Arab Spring, and the Black Lives Matter movement, they have also facilitated the spread of conspiracy theories, hate speech, and divisive rhetoric. Once relegated to fringe offline communities, these types of speech and ideas have now been mainstreamed. Even though research shows that only a small minority of users post such content, most people tend to believe that social media is flooded with hate and misinformation (Cato Institute, 2021). Does this raise the demand for more content moderation? What do citizens want from platforms?

In previous sections, we reported data on public perceptions of who should be responsible for maintaining a healthy online environment, and how people understand freedom of speech, as well as how normalized hate speech is across our country samples. In this section, we go a step further. We explore how citizens would want social media platforms to function if it were up to them. To do this, we use a simple trade-off. On the one hand, we presented participants with a hypothetical scenario of a platform where users can say whatever they want, with no content moderation at all, and on the other, we present them with a platform that is free of hateful speech and misinformation.

With freedom of expression re-entering the public debate and some platforms advocating for more absolutist forms of free speech (i.e., limited platform intervention), we sought to measure content moderation preferences within the context of this trade-off. We made the question abstract to gauge decontextualized public preferences. We encouraged respondents to think in terms of a scale that goes beyond the simple binary of moderation versus no moderation, viewing it instead as a continuous spectrum with shades of gray in between.

We asked respondents two separate questions; one about hate speech and one about misinformation. We first asked respondents to position themselves on a scale where, on one end, there are social media platforms with no content moderation whatsoever, and on the other, there

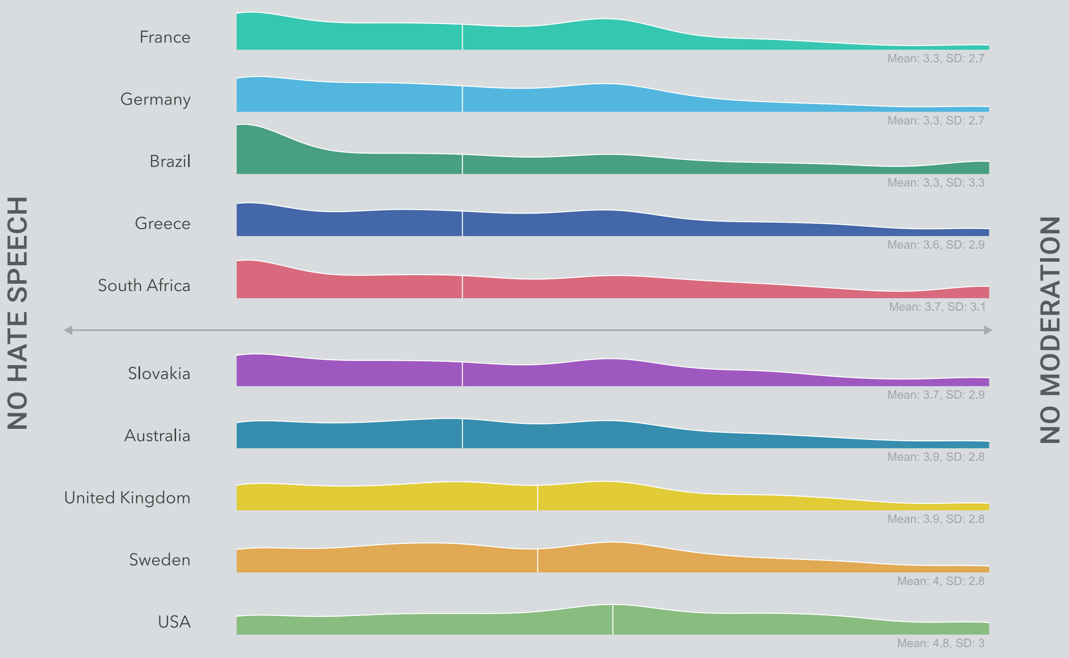


Figure 4.1 **Public Preferences: No Hate Speech vs. No Content Moderation**

"Imagine that there existed an option to have a social media platform completely free of hateful speech, or a platform where people can post whatever they want. On the scale below, what would be your preference?"

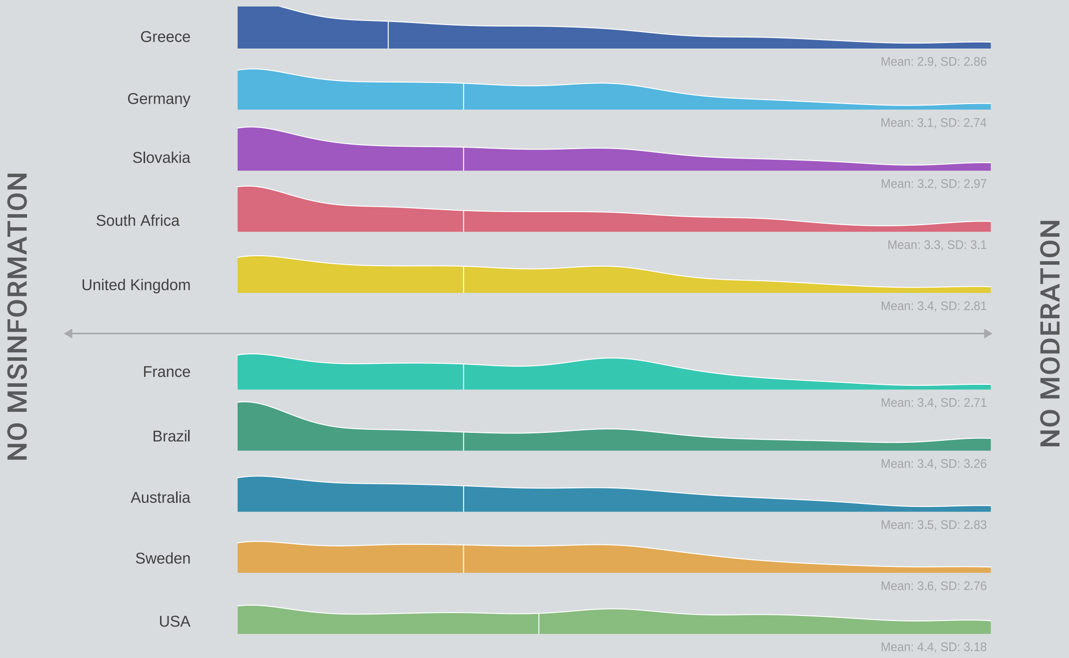


Figure 4.2 **Public Preferences: No Misinformation vs. No Content Moderation**

"Imagine that there existed an option to have a social media platform completely free of misinformation, or a platform where people can post whatever they want. On the scale below, what would be your preference?"

The vertical lines show the median, and the statistics show the mean and are platforms that exclude hate speech—described as intolerant, uncivil, or discriminatory posts. The results showed clear patterns, with most citizens around the world desiring some amount of moderation when it comes to hate speech. Even in the US, a country with a long-standing commitment to freedom of speech, it appears that, on average, Americans prefer some level of moderation to completely remove hate speech from social media. The countries that would require the most moderation to ameliorate hate speech were France, followed by Germany and Brazil. It should be noted that the differences among the top countries are relatively small.

Despite a strong commitment to free speech, even US respondents prefer some degree of moderation.

When respondents were asked to consider the same trade-off, but this time with content moderation aimed at producing a social media feed free from misinformation and fake news, the pattern was very similar. Most respondents valued online health more than the freedom to post whatever freely. The US remained an outlier, but again, most people would trade the freedom to post without restraint for a platform free of misinformation and disinformation. Across the board, it seems that respondents consider misinformation to be more important than hate speech when making this trade-off with Greece, Germany, and Slovakia having the lowest mean scores on the scale.

The percentage of respondents who choose one of the extreme options (9 or 10) for no content moderation varies across countries. The no moderation “absolutists” form a sizable group in the US (12.6%), Brazil (10.3%), and South Africa (9.3%), yet in the rest of the world, this figure is lower. Brazil and South Africa also have a relatively large percentage of respondents who desire protection from misinformation (37.1% and 38%, respectively), while the corresponding figure for the US is 24%.[1] Greece, as Figure 4.1 confirms, has the highest share of protectionists (41%) that, however, drops to 29% when hate speech is at the end of the trade-off. The same figures for Germany, France and South Africa are 32%, 32.5%, and 32.2%, respectively. The number of “absolutists”, finally, does not change substantially when the trade-off is between protection from hate speech and no content moderation at all.

Only in the U.S., Brazil, and South Africa do “no moderation absolutists” form a sizable group.

The takeover of Twitter by Elon Musk which was justified by allusions to an absolute freedom of speech has reignited the normative debate over the balance between freedom of expression and online content moderation. Since then, movements by both X and Meta toward a more laissez-faire approach have highlighted both the dynamic nature of platforms’ policies, and the difficult commercial, political, and ethical tradeoffs that these commercial players have to navigate. It is too early to judge whether recent moves will be beneficial in terms of user engagement and experience (and, ultimately, the bottom line). However, what we can say with some confidence is that most people do not want unmoderated platforms; they prefer some steps to be taken to reduce misinformation and hate speech in their feeds.

[1] These are respondents who chose options 0 or 1 on the scale.

REFERENCES

Baribi-Bartov, S., Swire-Thompson, B. and Grinberg, N., 2024. Supersharers of fake news on Twitter. Science, 384(6699), pp.979 – 982.

BBC, 2024. Men jailed for encouraging unrest on social media. https://www.bbc.com/news/articles/cy76dxkpjpjo

Beres, N.A., Frommel, J., Reid, E., Mandryk, R.L. and Klarkowski, M., 2021, May. Don't you know that you're toxic: Normalization of toxicity in online gaming. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1 – 15).

Bilewicz, M and Soral, W., Liu, J., 2020. Media of contempt: Social media consumption predicts normative acceptance of anti-Muslim hate speech and islamoprejudice. International Journal of Conflict and Violence (IJCV), 14, pp.1 – 13.

Bollinger, L.C. and Stone, G.R. eds., 2022. Social media, freedom of speech, and the future of our democracy. Oxford University Press.

Boulianne, S., 2018. Twenty Years of Digital Media Effects on Civic and Political Participation. Communication Research. 47(7), pp.947–966.

Brennan Center for Justice, 2021. Double standards in social media content moderation. New York University School of Law, pp.1 – 23.

Cato Institute, 2021. https://www.cato.org/survey-reports/poll-75-dont-trust-social-media-make-fair-content-moderation-decisions-60-want-more#

Das NETTZ, Gesellschaft für Medienpädagogik und Kommunikationskultur, HateAid und Neuen deutsche Medienmacher*innen als Teil des Kompetenznetzwerks gegen Hass im Netz (Hrsg.), 2024. Lauter Hass - leiser Rückzug. Wie Hass im Netz den demokratischen Diskurs bedroht. Ergebnisse einer repräsentativen Befragung. Berlin. https://komptenznetzwerk-hass-im-netz.de/download_lauterhass.php

Disinfo.eu, 2023. Sweden Disinformation Factsheet. Retrieved from https://www.disinfo.eu/wp-content/uploads/2023/05/Sweden_DisinfoFactsheet.pdf

Dreißigacker, A., Müller, P., Isenhardt, A. and Schemmel, J., 2024. Online hate speech victimization: consequences for victims' feelings of insecurity. Crime Science, 13(1), p.4.

Dvoskin, B., 2024. The Illusion of Inclusion: The False Promise of the New Governance Project for Content Moderation. Fordham Law Review, 2025, Available at SSRN: https://ssrn.com/abstract=4785833 or http://dx.doi.org/10.2139/ssrn.4785833

Dwoskin, E., Lorenz, T., Nix, N. and Menn, J., 2023. Antisemitism was rising online. Then Elon Musk's X supercharged it. https://www.washingtonpost.com/technology/2023/11/19/antisemiticism-Internet-elon-musk-israel-war/

Ellison, N.B. and Vitak, J., 2015. Social network site affordances and their relationship to social capital processes. The handbook of the psychology of communication technology, pp.203 – 227.

European Commission, 2024. The Digital Services Act Package. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Fletcher, R. and Nielsen, R.K., 2017. Are news audiences increasingly fragmented? A cross-national comparative analysis of cross-platform news audience fragmentation and duplication. Journal of communication, 67(4), pp.476 – 498.

Gillespie, T., 2018. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media. Yale University Press.

Gorwa, R., 2019. What is platform governance?. Information, communication & society, 22(6), pp.854 – 871.

Grierson, J., 2020. Surge in stalking victims seeking help during UK lockdown, The Guardian, 8 May. https://www.theguardian.com/uk-news/2020/may/08/coronavirus-surge-stalking-victims-seeking-help-during-uk-lockdown.

Grimmelmann, J., 2015. The virtues of moderation. Yale JL & Tech., 17.

Hate Aid, 2021. Boundless hate on the internet – Dramatic situation across Europe. https://hateaid.org/en/eu-survey-boundless-hate-on-the-internet/.

Heldt, A., 2019. Reading between the lines and the numbers: An analysis of the first NetzDG reports. Internet Policy Review, 8(2), pp.1 – 18.

Howard, J.W., 2021. Extreme speech, democratic deliberation, and social media. The Oxford Handbook of Digital Ethics, pp.1 – 22.

Kohl, U., 2022. Platform regulation of hate speech – a transatlantic speech compromise? Journal of Media Law, 14:1, 25 – 49, DOI: 10.1080/17577632.2022.2082520

Legal Resources Centre, 2020. A Critical Analysis of Content Moderation Policies and the Impact of Spreading Violence, Hatred & Disinformation in the Global South. https://lrc.org.za/wp-content/uploads/LRC-CONTENT-MODERATION-RESEARCH-REPORT.pdf

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S. and Hertwig, R., 2023. A systematic review of worldwide causal and correlational evidence on digital media and democracy. Nature human behaviour, 7(1), pp.74 – 101.

Mashable, 2024. X / Twitter use is down by nearly a quarter since the Musk Era started. https://mashable.com/article/x-twitter-daily-active-users-mobile-app-decline-report-x-disputes.

Meta, 2025. More speech and fewer mistakes, Meta, 7 January. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/#:~:text=We%20will%20allow%20more%20speech,it%20in%20their%20feeds%20can.

Müller, K. and Schwarz, C., 2020. From hashtag to hate crime: Twitter and anti-minority sentiment, conditionally accepted at AEJ: Applied Economics.

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson and C., Nielsen, R. K., 2021. Reuters Institute Digital News Report 2021. Reuters Institute for the Study of Journalism, Available at SSRN: https://ssrn.com/abstract=3873260

New York Times, 2017. The A.C.L.U. Needs to Rethink Free Speech. https://www.nytimes.com/2017/08/17/opinion/aclu-first-amendment-trump-charlottesville.html

New York Times, 2023. Whistle-Blower Says Facebook 'Chooses Profits Over Safety'. https://www.nytimes.com/2021/10/03/technology/whistle-blower-facebook-frances-haugen.html

New York Times, 2024. Book Bans Continue to Surge in Public Schools. https://www.nytimes.com/2024/04/16/books/book-bans-public-schools.html

New York Times, 2025. Meta's Decision to End Fact-Checking Could Have Disastrous Consequences. https://www.nytimes.com/2025/01/14/opinion/meta-fact-checking-policy.html

Pew Research Center, August 2020, Most Americans Think Social Media Sites Censor Political Viewpoints.

Pew Research Center, January 2021, The State of Online Harassment.

Pew Research Center, September, 2021, News Consumption Across Social Media in 2021.

Pew Research Center, September, 2023, Americans' Dismal Views of the Nation's Politics.

Pluta, A., Mazurek, J., Wojciechowski, J., Wolak, T., Soral, W. and Bilewicz, M., 2023. Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain. Scientific Reports, 13(1), p.4127.

Rathje, S., Van Bavel, J.J. and Van Der Linden, S., 2021. Out-group animosity drives engagement on social media. Proceedings of the National Academy of Sciences, 118(26), p.e2024292118.

Relia, K., Li, Z., Cook, S.H. and Chunara, R., 2019, July. Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 US cities. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 13, pp. 417 – 427).

Roberts, S.T., 2019. Behind the screen. Yale University Press.

Siddique, H., 2024. More than 300m children victims of online sexual abuse every year, The Guardian, 27 May. https://www.theguardian.com/society/article/2024/may/27/more-than-300m-children-victims-of-online-sexual-abuse-every-year.

Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., Stringhini, G. and Nilizadeh, S., 2023, July. SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. In 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P) (pp. 868 – 895). IEEE.

South African Human Rights Commission, 2017. Report of The National Hearing on Racism and Social Media in South Africa. https://www.sahrc.org.za/home/21/files/Racism%20and%20Social%20Media%20Report.pdf

Stevens, F., Nurse, J.R. and Arief, B., 2021. Cyber stalking, cyber harassment, and adult mental health: A systematic review. Cyberpsychology, Behavior, and Social Networking, 24(6), pp.367 – 376.

Švec, M., Madleňák, A., Hladíková, V. and Mészáros, P., 2024. Slovak Mimicry of Online Content Moderation on Digital Platforms as a Result of the Adoption of the European Digital Services Act. Media Literacy and Academic Research, pp. 81 – 95.

Taylor, J., 2023. X fined $610,500 in Australia first for failing to crack down on child sexual abuse material, The Guardian, 19 October. https://www.theguardian.com/technology/2023/oct/16/x-fined-610500-in-australia-first-for-failing-to-crack-down-on-child-sexual-abuse-material.

Tech Policy Press, 2024. We know a little about Meta's "Break glass" measures. we should know more. https://www.techpolicy.press/we-know-a-little-about-metas-break-glass-measures-we-should-know-more/.

Tucker, J.A., Theocharis, Y., Roberts, M.E. and Barberá, P., 2017. From liberation to turmoil: Social media and democracy. Journal of democracy, 28(4), pp.46 – 59.

Washington Post, 2023. Macron says social media could be blocked during riots, sparking furor. https://www.washingtonpost.com/world/2023/07/06/france-macron-social-media-block-riots/

Wasserman, H., 2020. Fake news from Africa: Panics, politics and paradigms. Journalism, 21(1), pp.3 – 16.

Williams, M.L., Burnap, P., Javed, A., Liu, H. and Ozalp, S., 2020. Hate in the machine: Anti-Black and Anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. The British Journal of Criminology, 60(1), pp.93 – 117.

Whitehouse, S., 2023. 'Section 230 Reforms,' in L. C. Bollinger and G. R. Stone (eds) Social Media, Freedom of Speech, and the Future of our Democracy. Oxford University Press.

Williams, Z. 2024. Racism, misogyny, lies: how did X become so full of hatred? And is it ethical to keep using it?, The Guardian, 5 September. https://www.theguardian.com/technology/article/2024/sep/05/racism-misogyny-lies-how-did-x-become-so-full-of-hatred-and-is-it-ethical-to-keep-using-it.

Wired, 2021. Facebook uses deceptive math to hide its hate speech problem. https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/

York, J.C., 2022. Silicon values: The future of free speech under surveillance capitalism. Verso Books.

YouGov, 2024. Support for under-16 social media ban soars to 77% among Australians. https://au.yougov.com/politics/articles/51000-support-for-under-16-social-media-ban-soars-to-77-among-australians.

CONTENT MODERATION LAB

Platforms, Users, and Free Speech

The Content Moderation lab, based at the TUM Think Tank, conducts empirical research on the supply and demand sides of online expression and platform content moderation.

On the supply side—that is, on those who provide the venues for online expression—we examine how digital platforms regulate content, their motivations, their stated moderation policies, and their actual practices. Given the significant differences in how various countries approach online speech (e.g., the contrasting regulatory landscapes of the U.S. and Europe), we explore platform governance comparatively within diverse legal frameworks and cultures of freedom of expression.

On the demand side—that is, the users who engage with platform-provided spaces, consume information, and create content—we investigate how platform decisions influence citizen behavior, public attitudes, and democratic participation. A core aspect of our research is the comparative study of public opinion on content moderation, censorship, and free speech, as well as how citizens perceive the role of social media companies in modern societies. Our goal is to uncover the mechanisms of platform governance and its broader societal impact, particularly in fostering democratic engagement rather than restricting it.

Through our research, we aim to provide insights that inform platforms, citizens, civil society actors, and policymakers. By generating evidence-based knowledge, we hope to contribute to more informed decision-making on regulatory matters, helping to shape policies that are democratically beneficial and conducive to healthier, more inclusive online spaces.

Our lab employs a range of methodological approaches, including experimental studies, cross-national comparative research, and digital trace data analysis. We collaborate closely with civil society organizations working on online regulation and problematic content, as well as with government institutions. Additionally, we actively engage junior scholars and students in empirical research on online expression and content moderation.

The lab is a joint initiative between the Technical University of Munich (TUM) and the University of Oxford, led by Professor Theocharis (Chair of Digital Governance, TUM) and Spyros Kosmidis (Associate professor, Oxford).

A report by the Content Moderation Lab
at the TUM Think Tank and the
Chair of Digital Governance

ACADEMIC INSTITUTIONS

Technical University of Munich
University of Oxford

RESEARCH TEAM

Yannis Theocharis (Technical University of Munich)
Spyros Kosmidis (University of Oxford)
Jan Zilinsky (Technical University of Munich)
Franziska Pradel (Technical University of Munich)
Friederike Quint (Technical University of Munich)

CONTACT EMAIL

contentmoderation.ttt@hfp.tum.de

PUBLISHER

Chair of Digital Governance and TUM Think Tank
at the Munich School of Politics and Public Policy
Technical University of Munich
Richard-Wagner-Str. 1 80333, Munich Germany

DESIGN & LAYOUT

Larissa Wunderlich

FUNDING & SUPPORT

This publication was funded by the Technical University of Munich and the University of Oxford.