HENRY M. JACKSON SCHOOL
OF INTERNATIONAL STUDIES

UNIVERSITY of WASHINGTON

Disputable Content and Democracy: Freedom of Expression in the Digital World

Task Force

The Donald C. Hellmann
Task Force Program

2023

The cover image: ATLAS Social Media (Photographer). (2017, February 10). Retrieved from https://www.flickr.com/photos/atlassocialmedia/32786886766.

Report font: The report font is OpenDyslexia Font. For more information about this font: https://opendyslexic.org/

Henry M. Jackson School of International Studies

University of Washington, Seattle

Task Force Report Winter 2023

Disputable Content and Democracy:

Freedom of Expression in the Digital World

Faculty Advisor

Dr. Jessica L. Beyer

Evaluator

Daniel Arnaudo

National Democratic Institute

Editors

Margaret Downey

Jasmine Pritikin

Project Manager

Isabel Wilson

Researchers

Victoria Chung

Christian Gales

Emily Glenn

Levi Howard

Tessa Kelly

Irene Lay

Avery Leonard

Sophie Maglaras

Gunhee Min

Thuy Nguyen

Jillian Ryan

Elena Van Eenenaam

Vienna Wang

Table of Contents

Executive Summary

Social media platforms offer a new medium to engage with others, express opinions and ideas, disseminate information, and interact with government officials. Nonetheless, social media has brought forth a new set of challenges, primarily the propagation of disputable content and its regulation. We define disputable content as content that is not explicitly illegal but can be understood as content whose accuracy or intention is questionable, misleading, untrue, or has the potential to incite violence, criminal behavior, or generally cause harm. In tackling disputable content, democracies find themselves at a crossroads. Some policymakers argue that tighter regulations are necessary to ensure the platforms' safety and protect users, while others perceive moderation as a violation of freedom of expression. Striking a balance between the two is difficult, and these complexities make effective policy and guideline creation increasingly challenging.

Our report looks at three major categories of disputable content, provides an overview of the current state of social media platform content moderation, and assesses content moderation legislation—or lack thereof—in democratic nations in order to understand the strengths and shortcomings of regulatory approaches to moderating social media content. To accomplish this, we unpack categories of disputable content with a particular focus on disinformation and misinformation, state sponsored content, hate speech, and extremist content and illustrate the real-world impacts of this content. We examine five social media platforms with significant global audiences—YouTube, Instagram, Facebook, TikTok, and Twitter—to understand their existing content moderation policies. Finally, we analyze government regulation related to social media content that exemplifies different approaches to this issue in democracies, including the European Union, Germany, the United States, Taiwan, Brazil, and Australia.

In tackling the issue of disputable social media content, this report offers democratic governments and social media companies eight recommendations. The following recommendations include four voluntary actions:

1. Institute a government-sponsored collaboration with international organizations and civil society to identify misinformation.
2. Encourage collaboration between social media platforms and the government in the development of content moderation policies beyond legal means.
3. Establish and disclose stricter policies for moderating leaders and elected officials online.
4. Prohibit the ability to purchase verification and establish a stricter verification process.

We also offer an additional four recommendations for greater legal regulation:

5. Implement clear definitions for disputable content.
6. Implement mandatory moderation transparency reports for social media companies.
7. Form a multinational body to establish a unified legal mandate requiring proportional content moderator employment.
8. Establish annual investment increases in AI moderation algorithm development to combat biases more efficiently.

Report Findings

This report provides a comprehensive overview of social media content moderation practices in democratic countries in order to examine the current landscape and identify areas for improvement moving forward. The recommendations presented in this report are based on a detailed analysis of types and impacts of disputable content, regulation of such content in six democratic countries, and the content moderation policies of five major social media platforms.

We find that one of the most pressing issues related to disputable content is the lack of universal definitions, including disinformation and misinformation, hate speech, and extremism. As a result, there is no unified understanding of what exactly to look for in the online space, which makes moderation particularly challenging. Without a baseline understanding of the defining characteristics of disputable content, it is difficult to take effective steps towards implementing successful moderation strategies.

Moreover, striking a balance between safeguarding user protection and preserving freedom of expression remains a significant challenge that requires the cooperation of various stakeholders, including governments, social media companies, civil society organizations, academic institutions, and think tanks. Although some companies and governments have made efforts to facilitate such collaborations, there is still much room for improvement.

Definitions

We define disputable content as content whose accuracy or intention is questionable, misleading, untrue, or has the potential to incite violence, criminal behavior, or generally cause harm. Furthermore, usually disputable content is not explicitly illegal in democratic countries, but the posting and dissemination of such content may lead to such activities. Under this definition, disputable content may include hate speech, misinformation, or activities by foreign actors that may encroach on domestic affairs. Conversely, disputable content does not include sexual exploitation or human trafficking, both of which are explicitly illegal activities.

Within our report, disinformation refers to intentionally misleading or false information that is spread with the purpose of deceiving or manipulating people; while misinformation refers to inaccurate or false information that is spread unintentionally or without the intention to deceive. Hate speech refers to any speech or behavior that expresses prejudice, discrimination, hostility, or violence towards individuals or groups based on their race, ethnicity, nationality, religion, gender, sexual orientation, or other characteristic. Finally, extremism refers to the holding of extreme political, religious, or ideological views that are seen as outside the mainstream.

It is, however, important to note that these definitions are not universally shared. Different stakeholders, including governments and companies, have their definitions or understanding of what these categories are and entail.

The Problem

Our report begins with an in-depth examination of disputable content: what that content is and what it looks like and the potential social harms disputable content may create. In particular, we focus on disinformation and misinformation, nation-state actors, hate speech, and extremism, topics that exemplify content that exists at the intersection of freedom of expression and societal harms.

These three specific types of disputable content were chosen because they are prevalent on social media today. In addition, their potential for causing harm to society and initiating real life consequences is substantial. Because of speech and expression rights, disputable content operates in a grey area for most government policies, leaving its mediation to the discretion of private social media platforms.

A key finding is that platforms struggle to operate at the intersection of freedom of expression and societal harms. Social media allows the discussion of various ideas, opinions, research, and points of view. In alignment with democratic values, most of the major platforms attempt to uphold values of free expression and speech on their platforms. The flipside of this is that this strong commitment allows for disputable content to proliferate. As online spaces are privately owned, platforms are able to moderate however they want in most democracies. When this is coupled with a lack of concrete definitions, if a post cannot be definitively labeled as violating a platform's community guidelines, companies opt

to keep that content posted to uphold a commitment to freedom of expression across their platforms.

Across our research, it is abundantly clear that this struggle to moderate and effectively take down disputable content has led to implications in the real world. Social media has the ability to incite violence, propagate terrorist content and recruitment, manipulate public opinion, influence elections, and spread false and dangerous information that may seriously impact politics, health, and general public safety. For example, former U.S. President Trump was able to leverage social media to a great extent during his time in office from 2016 - 2020, using platforms like Twitter to directly communicate with his supporters and the general public. During the 2020 presidential election, Trump made numerous false claims about voter fraud and irregularities, which he shared widely on social media, leading to increased distrust in U.S. elections. He also used these platforms to encourage his supporters to engage in acts of violence and unrest, most notably during the January 6th insurrection at the U.S. Capitol. Additionally, Trump's promotion of unproven or dangerous treatments for COVID-19, such as hydroxychloroquine, led to spikes in exposures to cleaners and disinfectants that put people's health at risk. All these instances put not only citizens at risk, but also long-standing democratic processes and confidence in the U.S. election system.

We also see this struggle come to head in the current *Google vs. Gonzalez*, a pending case in the United States Supreme Court. In this case, Reynaldo Gonzalez, the father of a victim of the 2015 Paris terrorist attacks, sued Google (specifically YouTube) for allegedly providing material support to the shooter and the greater terrorist Islamic State (Oyez, 2023). Particularly, Gonzalez alleges that Google provides assistance and support to international terrorism by permitting ISIS to exploit its platform for purposes such as recruiting members, planning terrorist attacks, making threats, instilling fear, and intimidating civilians, all facilitated through Google's computer algorithms that suggest content (Oyez, 2023). The Gonzalez case directly questions the role and liability that social media companies have for the content posted on their platforms. The outcome of this case could completely change the moderation landscape as it would impose that liability on the social media companies, companies that have long been protected in the United States under Section 230. Section 230 provides legal protection to online platforms such as social media websites, search engines, and other internet

intermediaries from liability for content posted by their users. *Google vs. Gonzales* is an exemplary example of real-life consequences that may occur without adequate moderation. The case also brings forth questions about platform liability, a primary controversy surrounding social media moderation.

In addition, algorithmic content sharing remains an issue when it comes to moderating disputable content. Algorithmic content sharing is a social media platform's use of algorithms to share new content with users based on that user's past behavior (GIFCT, 2021). On all platforms these algorithms typically collect information based on the user's previous behavior to recommend new but similar content (GIFCT, 2021). Algorithmically curated content presents a huge issue for moderation as interacting with disputable content means a person will be sent greater amounts of similar content, creating a dangerous cycle. For example, TikTok was found to further circulate hate speech content through algorithm-generated recommendations as a result of key words being substituted, while White supremacists' use of memes to successfully disguise, disseminate, and popularize disputable content across social media platforms has been identified as a significant issue. This is a problem across all types of content that exist on social media regardless of its topic: terrorist and alt-right content, political propaganda, health content like COVID-19 and vaccine misinformation, and much more. Not only do the methods disguising this disputable content make it difficult to identify and moderate, but the cultural and linguistic aspects of the content range drastically, a phenomenon which demands that moderating systems retain diverse knowledge on these nuances.

Private Sector Behavior

In choosing what social media platforms to examine, we selected platforms that operate around the world and reach large global audiences. YouTube, TikTok, Facebook, Twitter, and Instagram all rank within the top ten platforms in terms of the number of monthly active users worldwide (Walsh, 2022).

As mentioned above, when it comes to disputable content on their platforms, the decision to moderate or not is up to the platforms themselves—they are not necessarily bound by freedom of speech protections, although, it is certainly something these platforms take into consideration as they have large user bases that reside democracies where those protections are strongly valued. At the end of the day, social media companies are businesses, and as seen throughout this report,

typically prioritize keeping disputable content visible so long as it doesn't explicitly and clearly violate guidelines.

All the platforms have community guidelines that outline acceptable user conduct in order to promote a secure and enjoyable space. Each of the platform's community guidelines address our three main areas of disputable content, although how those companies refer to the disputable content differs by platform.

None of the social media companies examined in this report included the word 'disinformation' in their community guidelines. All platforms except for Twitter included the word 'misinformation' in their guidelines, although each one has a different approach to addressing and defining these concepts. Twitter instead categorized this type of content as synthetic and manipulated media, which refers to any media that can be misleading. This highlights the need for clear definitions that distinguish between terms because they have different implications and meanings.

Hate speech is outlined in the community guidelines of every social media platform that was examined for this report, although their policies for moderating this type of content vary. Parallel aspects of the definitions provided by these platforms were identified in the terminology, as each used the term "hate speech" directly, as well as the fact that most of them mentioned specific examples and/or subcategories that further clarify what constitutes hate speech on their platform. TikTok was the only platform which lacked any breakdown or expansion of their hate speech definition.

Similar regulations were found for extremism. The community guidelines for each platform all mention the term "extremism" directly and have those groups banned across all platforms. The general zero-tolerance consensus and disapproval of these groups represents the most unified position taken by social media companies out of all three areas of disputable content studied. Groups promoting ideas and recruiting for extremist purposes have developed tactics to avoid identification and violating guidelines, an occurrence that is seen on all five platforms. For instance, YouTube has experienced extremist groups taking advantage of their educational exceptions to content moderation to disseminate violative content, some which is further disguised to generate more views by reflecting mainstream trends. ISIS material was revealed to still be widespread on Instagram, and the reposting features on TikTok have helped circulate White supremacist and far-right memes. Currently all platforms' guidelines are insufficient due to a lack of specification and understanding of the diverse tactics employed by extremist groups.

In terms of moderation, all platforms use the same process for flagging content that potentially violates their guidelines. The first wave of moderation is done

by artificial intelligence, through the creation of machine learning algorithms that detect inappropriate or disputable content. From there, that flagged content is reviewed by a human reviewer who manually inspects it to decide whether the content breaches any regulations. Many of the same issues arise here—the inherent biases both engrained in algorithm creation and the actual moderators, problems with AI missing or incorrectly flagging disputable content, and insufficient preventative measures. In addition to these moderation mechanisms, users across all the platforms are able to anonymously flag content that they feel violates community guidelines or should be reviewed for any purpose. Like AI flagging, user-flagged content is sent to human moderators who review those posts to determine whether it violates community guidelines.

A notable commonality we observed across all five platforms was the use of third-party organizations for fact-checking. Meta's companies Instagram and Facebook both partner with third-party fact-checking firms which are verified by the International Fact-Checking Network (IFCN). Prior to Musk's takeover, Twitter similarly relied on reports from third parties such as PolitiFact, Lead Stories, SciVerify, Agence France-Presse, Animal Político, and Estadão Verifica to help determine when information was fabricated, although they relied more heavily on their own technology as a first line of defense in identifying the content (Perez, 2021). In contrast, YouTube and TikTok consult third parties rather than partner with them. Specifically, for certain types of content YouTube will provide fact-checking panels which link directly to reputable sources, such as the World Health Organization (WHO) for information panels related to COVID-19.

All the platforms claim that they are closely watching content that exists at the intersection of freedom of expression and societal harms. However, all platforms have drawn criticism for the actions they are actually taking. The primary criticisms include inconsistent enforcement of policies, such as the lack of enforcement for celebrities and politicians, and the reluctance to be transparent about any internal process including moderation criteria and process, human moderator demographics, and decisions about censorship. Despite claims to strong moderation tactics, it is clear that all five social media platforms must do better.

Government Regulation

In order to understand how governments reconcile the potential harms of disputable content with freedoms of expression, this report looks at six democracies: the European Union, Germany, the United States, Taiwan, Brazil, and Australia. To

decide which countries to examine, we first looked at democracies around the world to identify democratic governments' regulatory attempts to deal with content on social media platforms with a particular focus on countries that had influential or important content moderation laws. From that survey, we chose regulation that exemplified different approaches democracies are taking to deal with the issue of content that exists at the intersection of freedom of expression and societal harms.

Our research found that five of the six chosen countries have overarching laws that have been created or proposed, in part, to address the concerns that have risen with the boom of the internet and social media: the European Union's Digital Services Act (DSA), Germany's Network Enforcement Act (NetzDG), the United States' Section 230, Taiwan's Digital Intermediary Services Act (DISA), and Brazil's Marco Civil da Internet. Australia has the Broadcasting Services Act, which establishes a foundation for social media content moderation, although this regulation only outlines moderation of explicitly illegal content such as child exploitation images and terrorism. The DSA and NetzDG are the most comprehensive in terms of moderating disputable content and placing greater liability on social media companies—both have legally and non-legally binding aspects. The Marco Civil da Internet establishes an online legislative foundation, but lacks any liability, which creates gaps and difficulties when it comes to regulating content that is not explicitly illegal. Section 230 and the Broadcasting Services Act were both passed in the 1990s, so they both provide the foundation that modern social media moderation is built upon. Finally, Taiwan's DISA was introduced in 2022 and would both govern and outline the responsibilities of social media companies—including regular transparency reports and removal of violative content, similar to the DSA, although it has yet to be passed by Taiwan's Legislative Yuan.

When it comes to moderating disinformation and misinformation, Germany is the only country that has specific, legally binding legislation, not just for disinformation and misinformation, but for all three categories of disputable content we examined. Their Network Enforcement Act or NetzDG imposes fines on social media companies for non-compliance in the review and removal of posts. The fines effectively place financial liability on companies for removing illegal and harmful content, if a complaint is logged for a post that contains explicit content, that company has to take the post down within 24 hours; if the content is not explicitly illegal, the social media company then has seven days to investigate and decide whether or not to remove the post. The NetzDG also requires biannual complaint reports, which provide information about all posts both flagged and removed. These regulations and fines apply to all three categories discussed in this report.

In the European Union, legislators have opted to implement non-legally binding norms to combat disinformation and misinformation, as those categories are not explicitly mentioned in the legally binding DSA. Social media companies voluntarily sign on and agree to uphold those standards through implementing various methods of verification and fact-checking. The EU's Strengthened Code of Disinformation establishes a Code of Practice, a voluntary set of standards that agree to uphold. Like the requirements of Germany's NetzDG, this code mandates annual reports created by the social media companies to demonstrate their compliance.

Australia, Taiwan, Brazil, and the United States all have legislation pending that would create criminal and civil punishments for the dissemination of disinformation and misinformation—Australia's legislation mentions both disinformation and misinformation in its bill, Brazil only mentions disinformation, and the United States only mentions misinformation. These proposed regulations do, however, differ in their regulatory goals. For example, Brazil's Congress is currently debating a "Fake News" bill that specifically targets political disinformation, whereas in the United States there is a proposed amendment Section 230 focused on health misinformation. Taiwan's new proposed amendment the Presidential and Vice-Presidential Election and Recall Act and creation of the Special Act for Prevention, Relief and Revitalization Measures for Severe Pneumonia with Novel Pathogens seek to address both—however, its government has gone a step further by partnering the Taiwan FactCheck Center, which performs fact-checking of information related to public affairs, to further combat disinformation across the internet.

For hate speech, Germany and the EU are the only governments to have implemented specific legislation. Germany's NetzDG law regulates hate speech online in a parallel manner to content regarding disinformation and misinformation, whereas the EU has implemented both legally and non-legally binding policies for moderating hate speech. The EU's Council Framework Decision requires member states to institute laws which designate violative hateful content as a criminal offense, and their Code of Conduct on Countering Illegal Hate Speech Online is a voluntary agreement between platforms and the EU which encourages social media companies to take a more proactive approach on combating hate speech. Meanwhile, Taiwan, Brazil, Australia, and the U.S. have not created legislation to specifically target hate speech. Brazil and Australia have the Racism Law and Racial Discrimination Act, respectively, which prohibit racial discrimination, but only Australia's law definitively applies to online spaces.

Every country studied has regulation moderating extremism except for Taiwan. The only official legislation for this type of content in Germany is the NetzDG, which

once again aligns with their other approaches to disputable content such as hate speech and disinformation. Australia's Criminal Code Amendment includes provisions which indirectly target extremism and radicalization online, while the EU's Regulation on Preventing the Dissemination of Terrorist Content Online (TERREG) directly requires platforms to remove terrorist content within an hour of receiving a removal order from a competent authority. These three countries are the only ones to officially place some extent of liability on social media companies through their legislation on extremism, although the outcome of two pending court cases (*Gonzalez v. Google* and *Twitter v. Tammeh*) in the U.S. will determine whether the U.S. joins them. Although Brazil does not assign any liability to social media companies, their Anti-Terrorism Law criminalizes extremist activities online.

A common issue found across all the democracies we looked at was the challenges posed because social media companies are privately owned, with moderation being left to their discretion. For example, in the United States, the First Amendment to the Constitution guarantees freedom of speech as a fundamental right, but this serves as protection from the government, not private companies. Because social media companies are privately owned spaces, freedom of speech is not applicable to the content on those platforms. Within the U.S., this does not mean that individuals have an absolute right to say whatever they want on these platforms, and social media companies are generally free to set their own rules and guidelines for what content is allowed on their platforms, including restrictions on disputable content. They are not required to allow all forms of speech on their platforms, what they consider to violate their community guidelines are up to them.

This makes government regulation incredibly difficult, and because they are bound by the limitation of free speech—democratic governments cannot simply demand that platforms remove disputable content without facing strong criticisms of overstepping or potential legal rulings that a certain mandate is unconstitutional or violates freedom of speech protections. Because of this, governments are working to find a balance between protecting freedom of speech with the need to combat harmful content within the limitations of their legal rights. As we have seen, different governments have pushed their limits differently—Germany and the EU have taken a more uncompromising stance on the need to moderate disputable content, even in the face of criticisms related to freedom of speech. On the other hand, countries like the United States have struggled within their judiciary systems to find a balance.

Policy Recommendations

As social media users in democratic nations rely on major platforms for information and communication, democratic governments have a responsibility to support the advancement of content moderation online to make these spaces safer for the public. We recognize that there are many challenges with instituting large-scale legislation regarding content moderation, especially in terms of balancing freedom of expression rights and censorship concerns. Despite these obstacles, we believe that democratic governments have the resources and ability to make progressive impacts on the content moderation agenda for online spaces, an undertaking which certain nations are already partially doing.

With that in mind, we offer eight recommendations to democratic governments and social media companies. We provide four recommendations that can be implemented voluntarily, and another four that can be carried out through legal means.

The following recommendations include four voluntary actions:

1. Institute a government-sponsored collaboration with international organizations and civil society to identify misinformation.
2. Encourage collaboration between social media platforms and the government in the development of content moderation policies beyond legal means.
3. Establish and disclose stricter policies for moderating leaders and elected officials online.
4. Prohibit the ability to purchase verification and establish a stricter verification process.

We also offer an additional four recommendations for greater legal regulation:

5. Implement clear definitions for disputable content.
6. Implement mandatory moderation transparency reports for social media companies.
7. Form a multinational body to establish a unified legal mandate requiring proportional content moderator employment.
8. Establish annual investment increases in AI moderation algorithm development to combat biases more efficiently.

1: Institute a government-sponsored collaboration with international organizations and civil society to identify misinformation

We recommend instituting a government-sponsored collaboration that leverages the expertise of NGOs, academics, and other third-party organizations to identify and analyze misinformation. Taiwan's FactCheck Center (TFC) could be a loose model on which a partnership could be based, as the TFC similarly employs methods to coordinate with other fact-checking organizations around the world and analyze a diverse range of media to identify misinformation. But the TFC has many weaknesses, rooted in limitations in terms of resources, impact, reach, and language barriers. Since this multifaceted problem cannot be addressed effectively by one entity alone, partnering with a specialized network of civil society organizations would allow governments to effectively identify misinformation by working closely with experts. The TFC does include Communication scholars, fact-checking reporters, and lawyers, but governments should consider expanding their organizations to include broad expertise in areas such as media literacy and social media analysis.

Official protocols should be developed for sharing information and analysis among network members. These protocols should include a mechanism for sharing data, best practices, and emerging trends in misinformation. A key aspect of this initiative would be to solidify partnerships with at least one fact-checking organization based in each country. Forming a global lens for analysis is crucial to successfully identifying international trends for misinformation. Guidelines should also be created for collaboration between the network and democratic governments, including protocols for sharing information with government agencies responsible for public safety, national security, and election integrity. Government funding should be allocated to develop and maintain tools and platforms for analyzing and countering misinformation, as well as to provide training and support to network members. Eventually these networks could develop a unified worldwide collaboration, where efforts could become fully coordinated between individual democracies.

Although social media platforms have policies in place to address misinformation, our research revealed that inconsistent application of these moderation guidelines was prevalent. In some cases, even when presented with content which was fact-checked by third parties these companies were unsuccessful in enforcing regulation, exemplified by Facebook's handling of COVID-19 conspiracy theories. We also found platforms prioritized financial gains over user well-being, creating a barrier preventing platforms from upholding their policies. Therefore,

establishing specific government-sponsored networks to identify misinformation will provide clearer definitions of misinformation and disinformation, and hold social media companies accountable for identifying this type of content based on the more stringent criteria developed by the networks.

The secondary objective of this initiative would be to produce educational resources that would increase accessibility to reliable information for the public. Our research on social media case platforms illustrated that misinformation often comes from media outlets, especially in relation to content on COVID-19. Yet consistently potentially trustworthy information sources such as governments and professional experts attracted substantially less engagement across platforms, highlighting the need to bridge the gap and meet users where they are to ensure that these efforts to combat misinformation are successful.

Attention should be drawn to the government- and expert-sponsored educational content through marketing campaigns in order to help address the lack of awareness, and publishing clear reasoning and citations alongside fact-check findings will increase the credibility of the collaboration. These campaigns should target the spread of misinformation in areas such as health and politics, as our research on social media companies highlighted the disproportionate consequences resulting from the proliferation of falsehoods on these topics.

2: Encourage collaboration between social media platforms and the government in the development of content moderation policies beyond legal means

Establishing the legal precedent for moderating disputes is difficult, and the democracies we are looking at have encountered difficulty in passing legally binding policies due to controversies surrounding freedom of expression. While some actors, such as the EU and Germany, have been able to institute official legal measures, other countries struggle to pass policy through legislative bodies, effectively stalling the movement towards greater moderation.

To combat this issue, governments should encourage and initiate means of voluntary practices to hold social media platforms accountable for moderating content on their platforms. A strong example of what this could look like would be the European Union's Code of Conduct and Strengthened Code of Disinformation, where companies voluntarily agree to a self-regulatory set of standards. Other democracies should look to this as a starting point.

In the case of the EU, these codes work with the EU's larger and far more extensive Digital Services Act (DSA). Because of the extensive measures taken by the DSA to enhance content moderation, a framework such as the EU's Code of Conduct may not be possible given the lack of strong, comprehensive legislation elsewhere. Because the DSA incentivizes the voluntary participation of social media platforms in self-regulation, it is in the company's best interest to sign these codes as it ensures they are adhering to the strict DSA regulation. Therefore, if other democracies want to adopt similar measures, they must provide incentives for social media companies to voluntarily collaborate with other stakeholders.

Our case studies on social media companies indicate that these platforms focus on key priorities, including safeguarding their reputation, expanding their user base, maintaining positive investor relations, and maximizing profitability. By encouraging the collaboration between outside experts and stakeholders that work to make platforms as safe and accessible as possible for all people, social media platforms are able to establish trust with their investors and the communities they serve. To accomplish this, governments should partner with social media platforms to facilitate the establishment of a multi-stakeholder joint advisory committee that includes experts in content moderation, experts in disputable content such as the Southern Poverty Law Center, civil society representatives, NGOs like as the Anti-Defamation League, and other relevant stakeholders such as academics to provide guidance and stimulate dialogue on content moderation.

Platforms have already expressed interest in working together and being more involved in the conversation surrounding disputable content moderation, as can be seen with The Global Internet Forum to Counter Terrorism (Fioretti, 2017). Implementing this approach and formalizing collaboration for discussions surrounding hate speech, disinformation and misinformation, and extremism would ensure progress outside of formal legislation. Stakeholders could meet regularly to discuss and analyze emerging issues in voluntary agreements, as well as develop recommendations for improving existing systems. Ultimately, these conversations would seek to identify the best practices in content moderation beyond legal means and should eventually share ideas and recommendations across governments.

Joining these multi-stakeholder conversations would provide more credibility and accountability to social media companies regarding content moderation, as it would highlight their effort to continue to improve and stay up to date on disputable content discourse. Additionally, gaining access to the information and approaches utilized across governments would benefit all stakeholders and would inform the progression of content moderation initiatives internationally.

3: Establish and disclose stricter policies for moderating leaders and elected officials online

The research in this report shows that leaders and elected officials can be major distributors of both bad information online as well as create societal harm with that information. Social media companies should establish and disclose stricter policies for moderating leaders and elected officials online. As seen in the case of U.S. President Donald Trump and Brazilian President Jair Bolsonaro, elected individuals have a large unique opportunity to shape public opinion and their online statements may spur actions that can have real-world consequences. Stricter moderation policies for such individuals can help prevent the spread of harmful or false information that could threaten public health and safety or undermine democratic institutions. Due to the potential backlash that may arise from stricter guidelines, it is essential that moderation policies are made clear and imperative to promote greater accountability among public figures. While it may not be necessary to create an entirely new set of guidelines, social media companies should amend their current guidelines to include particular sections for elected officials. Doing so would be to the benefit of social media companies by building trust in social medial companies and assisting them in identifying appropriate ways to apply content moderation policies to political actors.

The definition of 'elected officials' should include any person democratically elected or appointed to a government position—from the president or prime minister all the way to town prosecutor or city governor. It is important that these guidelines continue to apply to elected officials after they have left office—influence doesn't disappear following the loss of an official title and it is imperative to continue recognizing the influence a past elected official still has over populations.

After establishing who falls under this category of 'elected official', social media companies could apply their existing community guidelines to these individuals, with additional scrutiny and enforcement measures. For example, if a leader or elected official violates community guidelines related to hate speech or disinformation, companies could implement more severe penalties than an ordinary user, such as a temporary or permanent account suspension depending upon the violation. In order to ensure transparency, public trust, and accountability in the application of these stricter guidelines, social media companies could also consider partnering with third-party organizations or experts in fields such as law or public policy from different political standings to provide further guidance and oversight.

These partners could help to ensure that decisions regarding the moderation of leaders and elected officials are fair, impartial, and consistent.

Brazil has taken a step in fortifying social media campaigning online within their election code. Political advertising on social media platforms is allowed only during a specific period leading up to the election. The Brazilian election code also includes provisions to prevent the spread of disinformation during election periods. While this stricter election code falls under the role of government, social media companies could take bits and pieces of this election code and include it in their community guides for greater regulation of elected officials. Twitter has taken a strong step forward here.

Ultimately, establishing and disclosing stricter policies for moderating leaders and elected officials online is critical for ensuring the integrity of online discourse and protecting democratic processes.

4: Prohibit the ability to purchase verification and establish a stricter verification process

Verification of accounts on social media are meant to confirm the authenticity and identity of an account on social media, which in turn establishes trust in platforms and creates a safer space. If any person is able to purchase verification for any reason, it completely undermines the credibility and legitimacy of both the verification process and the information on the platform itself. It is in the interest of social media companies to make verification processes support the spread of trustworthy information by authentic sources, rather than spread information by the highest bidder.

Tied to rejection of pay-for-verification policies, we recommend that companies create panels of people who represent different political orientations, as well as different backgrounds and lived experiences with a particular focus on those from marginalized groups to accommodate for cultural relativism, and experts in areas where there is significant disputable content such as medical misinformation.

Having a strict and cohesive process for verification helps prevent both fake or parody accounts, as well as bot accounts that may be used to engage in malicious activity such as spreading disinformation, promoting hate speech, or collective coordination in amplifying certain messages. By prohibiting the ability to purchase verification, the threat of manipulation of social media by bad actors is dramatically lowered. Much of the goal in moderating disputable content involves greater

transparency and accountability across all platforms, which the ability to purchase verification completely undermines.

The havoc that ensued following the roll out of 'Twitter Blue' shows the clear consequences of verification subscription. From an impersonation of pharmaceutical company Eli Lilly tweeting the company would provide free insulin to its customers to parody of former U.S. President George W. Bush accounts stating, "I miss killing Iraqis", it is abundantly clear that purchasing verification leads to an overall drop in platform trustworthiness (Mac et al, 2022; Corrons, 2022). The risks run anywhere from phishing attacks to geopolitical tension caused by impersonation of high-level officials (Corrons, 2022). Rather than learning from Twitter's experience, Meta has recently stated that it will implement a pay-to-verify policy for Facebook and Instagram (Capoot, 2023).

Social media accounts held by prominent figures have been found to receive less stringent moderation of content, and in some cases have avoided moderation or punitive action of any kind. This trend was identified in our social media case studies, and these gaps in moderation have most prominently allowed misinformation and hate speech to spread. Repeated negligence demonstrates that social media platforms have not established adequate ways to navigate the complexities of moderating users who received authentic verification status, indicating that they are not equipped to distinguish between paid and authentic verification in addition to standard accounts.

Pay-for-verification may also lead to less content moderation overall as platforms would be less incentivized to remove content if they can't clearly determine if a verified user who has paid is authentic. Not only will platforms lose credibility as paid verification erodes user trust, but the safety of users could be in jeopardy as the prevalence of disputable content would likely rise, as would its real-world consequences.

5: Implement clear definitions for disputable content

The lack of concrete definitions for disputable content is a significant issue for governments when enforcing and creating legislation for social media content moderation. Hence, democratic government should create and regulate clear definitions for disputable content such as disinformation and misinformation, hate speech, and extremism.

In every country we examined, governments had different definitions referring to disputable content. The EU, Australia, Brazil, and Taiwan, in particular, use vague definitions of disinformation and misinformation. Neither the United States, Australia,

Brazil, nor Taiwan have an official legal definition of hate speech or terrorism under federal law. The lack of specificity in definitions, and the lack of definitions altogether, make it challenging for governments to enforce regulation regarding the moderation of disputable content because the blurry criteria facilitates inconsistent and broad interpretations of these concepts that can serve as legal loopholes for social media companies. Instituting improved legislation is similarly difficult, as the lack of concrete definitions for this type of content creates ambiguity which allows for accusations of bias, censorship, or general encroachment on freedom of expression (Samples, 2019).

Thus, social media platforms have been left to moderate the disputable content on their platforms beyond government legislation, particularly in countries that have "safe harbor" regulations for their platforms, such as the U.S. Under these protections social media companies have little incentive to moderate beyond what is explicitly illegal or what they deem absolutely necessary. These challenges, coupled with the lack of concrete definitions for disputable content ultimately create a system with no effective way to enforce liability or moderation. Therefore, governments should establish clear and comprehensive definitions of disputable content accompanied by specific examples of the different ways in which the content may appear in online spaces.

In order to establish set definitions for disputable content, definitions should be written with the consideration that governments can justify their moderation of speech on social media if the regulations are narrowly focused and aimed at serving a "compelling government interest", a legal standard that courts use to evaluate whether a particular government action is constitutional (Samples, 2019). Essentially, if the government has a significant and legitimate reason for taking a particular action it can be enough to override other interests, including individual rights, such as freedom of speech (Miller, 2018). If the definition of a type of content is too vague or ambiguous, it could result in the government suppressing both allowed and prohibited speech, often ending with a decision of unconstitutionality.

In order to create these definitions, democracies should consult experts and civil society actors to ensure they are relying on research and data to inform their decisions about what to include. To ensure transparency, government actors should make the official definitions easily and publicly available, and they should provide insight to their processes and justifications for deciding on the definitions and subsequent regulation based on those established criteria. Finally, governments should develop a method for updating each disputable content definition on an annual basis, as it is crucial to evolve their regulation in accordance with the rapidly changing

digital world. This continual effort by democratic governments will ensure that concrete and accurate definitions remain intact, thereby preventing future issues of inconsistent and ineffective moderation, legal challenges, and unintended consequences for free speech and the protection of users.

6: Implement mandatory moderation transparency reports for social media companies

With the increasing importance of social media in shaping public opinion and influencing social behavior, it is crucial for social media companies to be transparent about their content moderation practices. To promote public safety and accountability, governments must take the lead in mandating annual transparency reports from social media companies. While every major platform already voluntarily provides quarterly transparency reports, they are very limited in terms of the content moderation information disclosed (Singh & Doty, 2021). We recommend that democratic governments should require transparency reports, as well as establish mandatory guidelines for social media companies to follow when creating these reports. These guidelines should demand detailed information on the number and types of content that have been flagged, removed, or allowed to remain on the platform. Additionally, the reports should provide insights into the company's content moderation teams, policies, and procedures, enabling the public to evaluate their effectiveness. Such reports would go a long way in enhancing transparency, ensuring user trust, and fostering public discourse. Hence, it is essential for social media companies to strengthen their transparency measures and be more accountable to their users and the public at large.

Other democracies should model these mandatory reports after the EU's Digital Service Act. Germany's NetzDG is similar legislation—in that it requires platforms to submit an additional transparency report, while Taiwan's Digital Intermediary Services Act (DISA) is currently pending implementation but would also require transparency reports on content moderation practices. However, the DSA is the best model as it requires annual transparency reporting and offers reporting requirements that establish a comprehensive, standardized, and effective way to promote transparency, accountability, and data protection on social media platforms. Modeling mandatory transparency reports after the DSA's requirements would ensure that social media companies are held accountable for their content moderation practices and to protect users' privacy and security. The DSA has an entire section on how large platforms must disclose all notices they receive, as well as the processes of identification of

specific disputable content. A key part of this transparency process is the availability of the data to academics, which allows for further analysis and understanding of the bulk data. It is clear from the adherence to the DSA's mandatory reporting that this is something social media companies are both willing and able to do. Establishing this kind of transparency across social media platforms in all parts of the world is a vital piece in ensuring the safety of users and preventing real-world harm.

However, there are criticisms that other democracies may need to address should they implement mandatory reporting. The first is that the DSA falls short in terms of tracking and reporting geographical distributions of material flagged. To combat this issue, governments of each country should outline requirements for platforms to sort their flagged material by region (according to the preferences of the government) and allocate funding for social media companies to partner with third-party data analytics organizations. Such standards will ensure equitable enforcement of the mandatory moderation reports as platforms are required to shift and adhere to the policy. The second major criticism comes from social media companies, who argue that the timeframes the EU provides for removing flagged content are unfeasibly short, as disputably illegal content is only given seven days for removal before a fine is issued. While it is important to ensure fair moderation practices, the feasibility of time allotments can be addressed by social media platforms hiring more moderators and improving AI moderation to better adhere to the legislation. More human moderators and better AI means that platforms will be able to assess the large quantity of flagged and reported content effectively within the timeframes provided.

To enhance the efficiency of content moderation on social media platforms, governments should convene a multi-stakeholder committee to discuss the logistics of expanding moderation teams. This committee should include experts from diverse fields, such as technology, law, human rights, and civil society. Through collaborative efforts, social media companies can gather resources and expertise to improve their hiring practices, secure funding from government or other sources, and receive consultation on other effective ways to moderate content within the given timeframe. By leveraging the expertise and support of all stakeholders, social media companies can ensure that their content moderation practices are aligned with the needs and concerns of governments and the broader public. Moreover, this approach would facilitate a productive relationship between governments and social media companies, leading to more transparent and accountable content moderation practices.

7: Form a multinational body to establish a unified legal mandate requiring proportional content moderator employment

We recommend that democratic governments collaborate to write and pass a unified multinational legal mandate for regulating content moderation on social media platforms. A multinational body of representatives from each democracy should be formed in order to draft this legislation to accommodate for laws and consider practices and preferences which currently exist across nations. The body would be responsible for setting content moderation standards and ensuring compliance by social media platforms, with the ultimate objectives being to promote online safety and reduce harmful content. Specifically, the moderation standards should outline the quantity of human moderators required for each respective platform and the punitive measures which will be enforced by the legislation passed across each country to ensure compliance by social media companies.

Human content moderators play a crucial role in ensuring that social media platforms are safe and conducive to healthy discourse. Investing in more moderators is necessary as one of the largest issues plaguing moderation is the sheer amount of information that needs to be reviewed. However, when investing in more moderators, companies must make increasing the number of moderators from different parts of the world a priority. This recommendation would require a certain number of content moderators for each democratic nation, thereby facilitating an increase in moderators who are able to understand various languages and cultural contexts. Moderators who are familiar with different countries, regions, cultures, and languages are more likely to understand the nuances of certain expressions or idioms and can better identify disputable content that violates community guidelines, which will also serve to improve the accuracy and effectiveness of social media companies' moderation efforts overall. This increase would also function as a counterweight to the biases ingrained in moderation, such as the cultural biases, linguistic biases, and political biases that result in certain groups being disproportionately impacted.

To determine the exact number of human moderators, the regulatory body should calculate criterion for each respective social media platform according to a minimum requirement for them to have at least one content moderator for every 10,000 active monthly users in a nation. To highlight why this quantity is needed we observe one of the largest human moderator workforces of Facebook and its Instagram subsidiary, which currently employs around 15,000 moderators, primarily through third-party vendors (Barrett, 2020). Considering Facebook alone, a significant

portion of their user base reside in the U.S., totaling 297.14 million monthly active users (McCain, 2022). With their current workforce, this would mean that each individual moderator would be proportionally responsible for reviewing violative content produced by approximately 19,809 users. Our report repeatedly confirms that this ratio is grossly inadequate, which is why we recommend the substantial expansion of human moderating teams for social media platforms to reflect at least one moderator per every 10,000 users. In the case of Facebook this ratio would essentially double their current workforce, and for other platforms could be an even larger increase. Thus, this ratio would ensure that social media platforms have an adequate number of moderators to identify and remove harmful content in a timely manner. Social media platforms that fail to meet this requirement would face penalties.

To enforce the requirements of social media platforms to have a proportional number of human moderators to monthly active users in a country, the regulatory body should implement fines. Creating fines and specific violation criteria will allow member countries to be transparent about the consequences of platforms failing to comply. A three-strike rule should also be instituted to demonstrate that if social media companies repeatedly violate, they will face a temporary ban from operating in a particular country. We recommend a tiered punitive action structure, which should outline strike one as a certain fine, strike two as a larger fine, and strike three with another substantial fine alongside a temporary ban from operating in the country where the violation occurred. Money generated from the fines should be pooled together for the multinational body to then allocate to various content moderation improvement initiatives as they see fit, such as funding enforcement efforts of legislation violations.

By implementing this policy recommendation, social media platforms can improve their content moderation practices, both in terms of effectiveness and cultural and linguistic sensitivity. This will help ensure that the platforms are adjusting their human moderation employment to more properly account for the substantial size of their user base, as well as for the diverse cultural norms present in each member country that they are used.

8: Establish annual investment increases in AI moderation algorithm development to combat biases more efficiently.

Platforms should annually increase their AI technology investment by 2%, a spending level which would be focused on developing more sophisticated moderation tools that can recognize and flag content that may be culturally insensitive or

inappropriate. Official enforcement of this spending increase should be achieved by government actors passing legislation in each individual country to require social media companies to actively improve their software moderation technology in this way. Drafting a unified law for democratic countries should be accomplished in a parallel manner to the approach described in Recommendation 5, as government representatives should form a multinational body to consider practices and preferences which currently exist across the nations. In this case, the body would be responsible for writing the law and supporting enforcement efforts for legislation violations, as well as developing a detailed format for social media company financial reports and guidelines for the spending increase. Platforms should have to submit compulsory annual financial reports which detail how their AI moderation budget is spent, particularly accommodating for the 2% increase, thereby ensuring accountability that the companies are utilizing their budget in accordance with the established guidelines. This legislation should include measures to fine social media platforms 10% of their annual profits should they not raise their spending on AI technology development by 2% each year and adhere to spending guidelines.

Guidelines for spending expectations should be created by the multinational body, although they should consult with third-party organizations and experts on progressive AI software development strategies and spending. We recommend that the guidelines should outline expectations for two primary areas. First, all improved tools should be based on machine learning algorithms that are trained on data from different regions and languages. Without diverse representation among developers and data sets, AI moderation algorithms can bolster or amplify existing biases because AI algorithms can be influenced by the biases of their developers and the data they are trained on. By promoting diversity in the development of AI moderation, algorithms are able to combat biases more efficiently. Because AI serves as the first wave of moderation for all social media, platforms must ensure that there is ample diversity within AI systems to combat racial, ethnic, gender, and socio-economic biases which may pervade their algorithms.

The second area we recommend outlining should support spending on progressive AI moderation. Current AI systems for major platforms are ineffective in many ways, demonstrating the need for new initiatives. To address these issues, the guidelines should support social media company spending on the pioneering of new initiatives, research, and approaches. Whether or not the innovative solutions are effective should not be part of the criteria, as positive and negative outcomes will both serve to inform AI moderation improvement regarding disputable content. An option for innovative reinvestment by platforms could be to create units like Google

Jigsaw. This think-tank unit of Google aims to develop technology for better combating problems relating to digital censorship, online harassment, and violent extremism. Progressive AI content moderation efforts through new initiatives, research, and targeted reinvestment strategies are essential to the continual improvement of AI technology and promoting a safer online community.

Disinformation and Misinformation

Tactics to spread disinformation and misinformation are evolving faster than social media platforms and government policies established to combat them. The distinction between trustworthy information and disinformation or misinformation on social media is critical to the integrity of democracies, fostering healthy online communities, and discerning accurate information. As such, it is an extremely important issue for social media platforms and their users.

The following section will discuss disinformation and misinformation spread in two ways. First, through foreign state-sponsored behaviors with the case studies of Russia in the United States and China in Taiwan, and second, through domestic actors with the case studies of President Donald Trump in the 2020 U.S. presidential election, and regarding the COVID-19 pandemic.

State-sponsored Behavior

As social media use has become ubiquitous, governments have increasingly used social media for the dissemination of disinformation and misinformation. While authoritarian regimes target their own populations with propaganda, they also target foreign publics (Bradshaw & Howard, 2017). Both Russia and the People's Republic of China (PRC) attempt to manipulate public opinion in democratic countries via popular social media platforms, often employing sophisticated covert tactics to maximize exposure and minimize suspicion of foreign influence. The consequences of these online campaigns are substantial. Successful campaigns may result in outcomes such as the manipulation of public opinion regarding political decisions abroad.

Russian Disinformation

Following a two-year investigation, the Justice Department special counsel Robert Mueller indicted 13 Russians and three Russian entities in 2018, as well as the so-called "troll farm" company the Internet Research Agency for attempted interference in the 2016 election (Benner & LaFraniere, 2020). Prosecutors charged the defendants with "conspiracy to defraud the United States," using social media to spread disinformation, attempting to subvert the 2016 election, and polarizing American voters. In 2020, the charges were ultimately dropped, although the efforts of Russian actors cannot be ignored (Benner & LaFraniere, 2020). Given the high exposure of the Internet Research Agency and their activities in the United States,

the organization presents a relevant and prominent case study on the tactics, motives, and effects of foreign social media campaigns in democracies.

There were several goals in Russia's activities through their social media campaigns: (1) polarize political beliefs among Americans, including targeting content to citizens based on race and party leanings, (2) promote voter suppression operations, (3) promote secessionist and insurrectionist sentiments, (4) engage in pro-Trump (and necessarily, anti-Clinton or Biden) campaigns, and (5) increase or erode support for prominent figures relevant to Russian interests, such as Robert Mueller (DiResta et al., 2019). To accomplish these goals, the Internet Research Agency employed a variety of tactics in their attempt to interfere in the 2016 and 2020 presidential elections in the United States.

One tactic the Internet Research Agency used was to target many different groups across political, racial, and religious spectrums. By posing as organic users, the Internet Research Agency would gain a following on their accounts targeted towards certain interest groups such as women, the Black people, and veterans (DiResta et al., 2019). On the political spectrum, right-targeted content promoted themes of anger and suspicion towards voter fraud, illegal participation in the election, or Clinton potentially stealing the election (DiResta et al., 2019). Left-targeted content was more anti-establishment and promoted voting for any candidate other than Clinton (DiResta et al., 2019). In its attempt to spread content which would confirm existing biases, the goals of the Internet Research Agency were clearly to further polarize Americans in their beliefs based on already present divisions.

In addition to exploiting the divisions based on group identity, Russian actors also used ads to target some of the most divisive issues in the United States (Collins, 2018). In 2018, an investigation by the House Intelligence Committee released thousands of paid Russian ads that were revealed to be specifically targeted towards Facebook users by age, geographical region, and interest (Collins, 2018). Over two-thirds of the ads were found to be related to race, while the remaining third predominantly concerned nationalism, immigration, and terrorism (Kim et al., 2018). The content of the ads was often framed in an us-versus-them distinction, promoting an in-group and demonizing an out-group (Kim et al., 2018). A closer look also revealed that there were high levels of online material being geographically targeted towards traditional battleground states during the 2016 election which promoted information on certain political issues (Kim et al., 2018). Due to uncertainty over how battleground states would ultimately vote in the election, these efforts indicate that Russia's motives may not only be to stimulate further political polarization of far-right or far-left groups, but also influence the entire election outcome. Observed in both the 2016

and 2020 U.S. presidential elections, these tactics were used to try and sway undecided voters towards casting their vote for Donald Trump through targeted anti-Clinton, anti-Biden, and pro-Trump ads (Howard et al., 2019).

Although the evidence clearly supports that there was Russian interference in the 2016 and 2020 elections, research remains inconclusive as to whether campaigns by foreign nation-states substantially impacted the voting behaviors, polarization, or political attitudes of Americans (Bail et al., 2019). However, it would be inaccurate to conclude that there was no impact altogether (Bail et al., 2019). News about Russian interference has raised questions about the legitimacy of elections, increased mistrust in the electoral system, and changed American willingness to accept claims of voter fraud in the current election or future ones (Devitt, 2023). In other words, the attempt by foreign nation-states alone is enough to sow seeds of doubt in the ability of social media platforms in democracies to adequately monitor disputed content.

Already, the example of the Internet Research Agency's social media campaigns in the United States have proven to be widespread. Data indicates that between 2015-2018 alone they reached 126 million people on Facebook, 20 million users on Instagram, 1.4 million users on Twitter, and posted over 1,000 videos on YouTube (DiResta et al., 2019). In appropriating the identities of real domestic nonprofit organizations or political action committees, the Internet Research Agency showed that it could simultaneously bypass tech platform policies that work against political campaigns by foreign actors and increase user engagement with its content (Kim, 2020). The lack of punishment against foreign influences in the 2016 election cycle may have inadvertently encouraged the Internet Research Agency and other foreign actors to continue their media operations (Kim, 2020). Evidently, should democratic governments and social media platforms fail to adequately address foreign state-sponsored manipulation in their politics, their tactics will continue to evolve and become more advanced.

Russia's disinformation and misinformation campaigns in the United States were not an isolated incident. A recent 2022 press release from the United Kingdom indicates that Russia's "cyber soldiers" have scaled up their operations to target other foreign democratic publics, such as in South Africa and India. Their focus has shifted to recent events, most prominently the war in Ukraine (Foreign, Commonwealth & Development Office et al., 2022). Tactics include spreading disinformation about the activities in Ukraine and spamming the social media accounts of Kremlin-critics and world leaders with pro-Putin and pro-war comments, attempting to widely disseminate and generate increased sympathizer sentiment for the illegitimate invasion (Foreign, Commonwealth & Development Office et al., 2022). To bypass detection by social

media platforms, Russian actors have also avoided making original content, instead focusing on commenting with VPNs (Virtual Private Networks) to amplify pseudo-"organic" content (Foreign, Commonwealth & Development Office et al., 2022). VPNs are a type of network connection that provides users with a secure and private connection over a public or shared network. Russia took advantage of the fact that this process hides the user's IP address and location, making it difficult to track their online activity. Analysis also revealed another evasion approach was accomplished by paying TikTok influencers to promote pro-Kremlin narratives (Foreign, Commonwealth & Development Office et al., 2022). Tactics such as these make it much more difficult for social media platforms to track and moderate disinformation campaigns by foreign actors ("'Troll factory'", 2022).

The tactics that the Internet Research Agency and other Russian agencies continue to use were clearly meant to interfere within the domestic affairs of the United States using social media, and the effects of this manipulation are tangible. To some degree, foreign nation-states prove to be able to polarize the ideological spectrum in America, manipulate public opinion towards their own agenda, and call into question the legitimacy of political institutions and politicians by undermining democratic processes. This demonstrates the complexities of identifying the perpetrators of disinformation and misinformation on social media, determining what content should be moderated, and considering the consequences of pursuing legal action against any offending foreign nation-state actors given current political relations.

Chinese Disinformation

As with Russia, the People's Republic of China (PRC) have also used social media campaigns in other democracies in an attempt to manipulate public opinion abroad. Most prominently, PRC disinformation and misinformation campaigns target Taiwan, further complicating long-standing, tense cross-Strait relations.

PRC strategies share many similarities with Russian actors, such as the Internet Research Agency. Chinese content farms have been used on multiple platforms, such as purchased Facebook accounts, text-messaging campaigns, networks of automated bots programmed to overwhelm Twitter hashtags with false information, and coordinated activities intended to manipulate search results (Lührmann et al., 2019). Although it is well-documented that these campaigns appear to support the political preferences of the Chinese Communist Party (CCP), due to the anonymity and complexity of these tactics, it is difficult to directly attribute it as state-sponsored

behavior (Lührmann et al., 2019). However, given the ambiguity of its origins, it may give the impression that China is behind every instance of disinformation, which serves to boost its image as a dominant power over an outmatched Taiwan (Harold et al., 2021). Overall, the content of these disinformation campaigns is to vilify those who oppose the CCP regime, such as President Tsai's administration, who is known to take an anti-unification stance (Quirk 2021). For example, messages may include inaccurate claims of the effects of the Tsai administration's policies, that the Taiwanese military is weak and corrupt, or that Taiwan serves as an example of why democracies are an "ineffective" and "chaotic" political system (Harold et al., 2021).

Knowing that they will likely not change the minds of enough Taiwanese to support unification of the mainland, campaigns to spread disinformation are not necessarily to sway public opinions towards a united China (Quirk, 2021). Instead, exploiting existing divisions between the Taiwanese people are meant to undermine the public image of the Taiwanese government as a healthy, functioning state and target moderates who are more susceptible to a pro-Beijing stance (Quirk 2021). As with the United States and other democracies, the case of the PRC campaigns in Taiwan show how polarization via foreign social media influence can be used to undermine the integrity of a democracy.

While PRC campaigns have long been used in an attempt to interfere with Taiwan's elections, like Russia in the United States, they were also used to spread disinformation and misinformation about the origins of COVID-19 and Taiwan's response to the pandemic, including the claim that Taiwan was covering up virus deaths or that President Tsai contracted the disease (Frenkel et al., 2020). Social media campaigns by China regarding COVID-19 have led to concerns that the PRC was using the pandemic as another tool to undermine Taiwanese democracy, a belief also held by John Wu, Taiwan's foreign minister (Frenkel et al., 2020). Much of the campaign work has been linked to China's United Front Work Department and the Communist Youth League, according to the U.S.-based Center for Strategic and International Studies ("Fake news", 2021).

As tactics and motives of foreign nation-states continue to evolve, it is up to governing bodies and social media platforms to keep pace. Although the privacy and anonymity of social media make it more difficult to monitor if foreign influence is involved among genuine user activity, the distinction of foreign influence from genuine social media use is critical to the integrity of politics in democracies, fostering healthy online communities, and discerning accurate representations of a nation's constituents. Hence, the identification of disputed propaganda content is intensely vital as social

media remains a primary source of information and communication, despite increasing difficulties.

By using tactics such as flooding online spaces with bot-produced content or misleading paid promotions, foreign state-sponsored social media campaigns have shown that countries like Russia and China are able to exploit the weaknesses among government and social media platform policies for their own gain in democracies around the world. The spread of disinformation and misinformation on social media has been observed to influence the opinions of individual users to proliferate distrust democratic institutions, cause further division and polarization within democratic societies, and conceal content that portrays foreign nation-states in a negative light, while simultaneously promoting positive and potentially misleading narratives. As social media manipulation tactics continue to evolve according to the agendas of foreign nation-state actors, democratic states will continue to be vulnerable to this type of content unless governments and social media platforms correspondingly evolve their moderation tactics.

Domestic Actors

The prevalence of misinformation produced by domestic actors on social media platforms has become a significant issue for social media moderators. The following section will discuss misinformation spread by domestic actors through an examination of President Trump in the context of the 2020 presidential election in the United States, and in regard to the COVID-19 pandemic. Three of the categories of misinformation from domestic actors that will be discussed are political, violence-inciting, and health misinformation. The spread of these types of content can have severe real-world consequences, including political divisions, loss of confidence in democratic systems, and public health risks. Therefore, effective moderation of such content is particularly crucial for protecting democratic institutions, public safety, and health.

Political Misinformation

Political misinformation has impactful real-world consequences, such as affecting the public's political opinions, reducing confidence in democratic voting systems (Ratliff, 2021), and creating political divisions and tensions (Quinnipiac University, 2022). The moderation of political misinformation is crucial in combating these consequences and ensuring that political processes are not negatively

impacted, which is why it has been a central issue for many social media platforms. The problem content moderators face regarding posts that potentially contain political misinformation is how to distinguish between varying opinions and potentially harmful content.

Political misinformation had a major impact on U.S. presidential elections and has created lasting problems in social media moderation regarding election integrity. Following the 2020 election, Trump was one of the largest sources of misinformation as he promoted false claims of election fraud both on social media platforms and in person (Sanderson et al., 2021). Known as his "big lie", Trump alleged that there were left-wing efforts to conspire against him and rig the election outcome in their favor. Ultimately, these claims amplified Russian efforts to undermine trust in America's democratic system and had lasting impacts on American voters' trust in election outcomes (Wolf, 2021).

One of Trump's main tools used for the dissemination of 2020 presidential election misinformation was Twitter, and the shortcomings of Twitter's content moderation in terms of removing and addressing potentially harmful tweets made by Trump facilitated his impact on public opinion during and after the election (Atske, 2021). In November 2020, Twitter released a statement saying that they had flagged approximately 300,000 election-related tweets as "disputed and potentially misleading" (Gadde & Beykpour, 2020). A news analysis conducted by Issue One, designed to document the spread of election misinformation, found that 60% of Trump's 1,500-plus tweets between the day after November 3, 2020, and January 8, 2021, when Twitter permanently suspended his account, were messages that sought to challenge the results of the 2020 election—an average of about 14 tweets per day. Following the 2020 election Trump falsely called the election "rigged" 57 times, alleged that the election was "stolen" or that the Democrats tried to "steal" it 57 times, inaccurately declared that he "won" the presidential election 25 times, and falsely called the outcome a "landslide" victory in his favor 12 times (Ratliff, 2021). These statistics raise the question of why President Trump's posts were not more closely moderated, or why his account was not removed earlier than it was. The unsuccessful efforts of Twitter's content moderators allowed misinformation spread by Trump to further circulate and potentially influence user opinion.

A CNN poll conducted in the summer of 2021 found that 36% of Americans do not believe that President Biden legitimately won the 2020 U.S. election, and among Republicans specifically that number leaps to 78% (SQL, 2021). An NPR poll conducted in October of 2021 found that 75% of Republicans say Trump has a legitimate claim that there were "real cases of fraud that changed the results" (Marist Poll, 2021).

These polls show the extent to which Trump's claims have impacted public perception of the 2020 election and of the American democratic system. This is a prime example of political misinformation that was not adequately moderated by social media platforms, which led to a lack of confidence in the American democratic system (Klar, 2022).

Additional studies found that the events and social media misinformation surrounding the 2020 U.S. presidential election led to distrust in the American democratic system among Americans. NPR conducted a survey in January of 2022 which found that 64% of the American population believes that U.S. democracy is in crisis and is at risk of failing (Ipsos, 2022). Similarly, a survey also conducted in January of 2022 by Quinnipiac University revealed that 76% of respondents thought political instability within the country was a bigger danger to the United States than external adversaries (Quinnipiac University, 2022). This poll also found that 53% of Americans expected political divisions in the country to worsen over their lifetime (rather than get better). Hence, these increases in political divisions and tension demonstrate some of the tangible problems which can arise from domestic misinformation campaigns aimed at disrupting U.S. democratic systems. It is crucial for social media platforms to develop effective moderation policies to prevent the spread of political misinformation, as this will ensure that users are protected against misleading content and the democratic process is not negatively impacted.

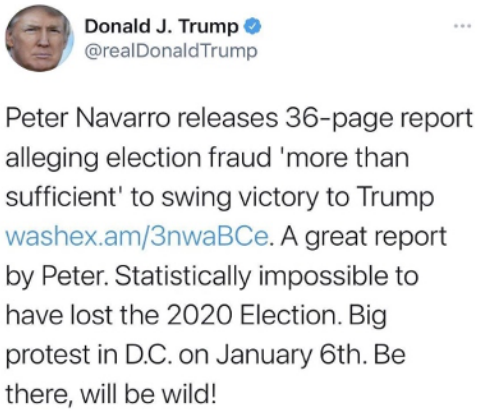
Violence-inciting Misinformation

Social media has also created a space where individuals attempt to use platforms to incite or promote violent acts, and misinformation posted with this intent is extremely dangerous due to the possibility of it having tangible violent outcomes. Thus, mentions of violence are of major concern when it comes to making sure this type of content is moderated by social media companies. Many major platforms strictly forbid content that calls for or has the potential to mobilize violent or criminal actions (Meta, 2023). However, content moderators still struggle to differentiate between innocuous posts and content that could be capable of provoking real-world outcomes yet doesn't explicitly promote harmful activity.

The phenomenon of misinformation on social media platforms fueling hate and leading to real-world violence is demonstrated in the January 6th attack on the capitol, as President Trump's tweets went further than just promoting his conspiracy of election fraud. On January 6, 2021, a mob of Trump's supporters attacked the U.S. capitol building in an attempt to prevent the joint session of Congress from counting

the electoral college votes that would formalize the victory of President-elect Joe Biden. The attack left five dead and hundreds injured (Rubin et al., 2022). Since the attack, 725 arrests have been made in relation to actions taken by individuals at the capital (Office of, 2021). A former Twitter employee who worked on content moderation policy argued that Trump's tweets inspired the violence at the January 6th insurrection. The tweet credited with inspiring this event was from December 19th, where Trump asked his supporters to attend a "Big protest in D.C. on January 6th" in response to an "impossible" loss of the 2020 election (see Figure 1).

Figure 1: President Trump Twitter Post about January 6th



Twitter banned President Trump from the platform two days after the attack, stating that the action was necessary "due to the risk of further incitement of violence." The platform cited two tweets posted in the days following the event to corroborate their decision, although many argue he should have been banned long before (Twitter, 2021). Twitter also confirmed that plans for upcoming protests involving weapons started to spread both on and off Twitter. These plans included a suggested additional assault on the U.S. Capitol and state capitol buildings on January 17, 2021 (Jones & Bey, 2021).

Although Trump was banned from Facebook and Twitter, many of his supporters have established deeply intertwined networks of online communities that continue to encourage future conflict and sow doubt in the democratic process (Heilweil & Ghaffary, 2021). An article written by Rebecca Heilweil and Shirin Ghaffary states that online extremism has spilled over into real-world consequences and that the January 6th attack on the capitol is undeniable evidence that online misinformation promoted by Trump has had genuine impacts on U.S. politics (Heilweil & Ghaffary, 2021). Hence, Trump's tweets are examples of how social media

misinformation, and the absence of effective regulation can lead to tangible consequences, in this case violence, underscoring the importance of social media moderation.

Health Misinformation

In addition to influencing the political sphere, misinformation had a considerable impact on the COVID-19 pandemic. Social media played a unique role in the COVID-19 pandemic, with the dissemination of health misinformation leading to what many are calling an "infodemic". According to the World Health Organization (WHO), an infodemic is "too much information including false or misleading information in digital and physical environments during a disease outbreak" (WHO, 2023). Social media created an avenue for people outside of the medical field to spread information about the COVID-19 outbreak, creating a challenge for content moderators attempting to debunk misinformation on social media, (Mourali & Drake, 2022). This has led to the spread of medical falsehoods, referred to by the WHO as health misinformation which is defined as "any health-related claim of fact that is false based on current scientific consensus" (Chou et al., 2021). President Trump serves as another example of content, this time in the form of health misinformation.

Health misinformation spread on social media can lead to real-world public health concerns (Lovelace Jr., 2020). In addition to spreading misinformation about election fraud, Trump also spread falsities regarding the COVID-19 pandemic. A study conducted by researchers at Cornell University found that Trump was the single largest driver of misinformation around COVID-19 (Evanega et al., 2021). The study analyzed 38 million articles about the pandemic in English-language media around the world. President Trump made up nearly 38 percent of the overall "misinformation conversation," making him the largest driver of the "infodemic" (Evanega et al., 2021). According to Dr. Joshua Sharfstein, a vice dean at the Johns Hopkins Bloomberg School of Public Health and a former principal deputy commissioner at the Food and Drug Administration, misinformation regarding the COVID-19 was "one of the major reasons" the United States struggled to combat the pandemic (Stolberg & Weiland, 2020).

Health misinformation can also lead to individuals engaging in self-harming activities (Lovelace Jr., 2020). The Cornell University study also looked at misinformation media coverage on the topic of 'miracle cures' in relation to COVID-19. They found that spikes in media coverage corresponded with misinformation spread by President Trump (Evanega et al., 2021). Early peaks in articles written on this topic

correspond with when President Trump began to advocate for the use of hydroxychloroquine and chloroquine as treatments or cures for COVID-19, even although there was no peer-reviewed clinical data to back up his claims of efficacy (Lovelace Jr., 2020) (Wang, 2020). This triggered a substantial spike in media coverage regarding his announcement, caused the subsequent shortage of these drugs, and later the finding that they were not effective in treating COVID-19 and might indeed be harmful, (Boseley, 2020). The largest and most notable peak corresponds with President Trump's press conference statements about the potential of injecting bleach or other disinfectants internally as a cure for coronavirus, (BBC, 2020).

Centers for Disease Control and Prevention reported that spikes in exposures to cleaners and disinfectants reported to U.S. poison centers corresponded with Trump's comments on injection disinfectants as a cure for COVID-19 (CDC, 2020). This is yet another example of how Trump's comments and subsequent spread of misinformation led to negative outcomes in the real world. Trump's use of Twitter and other social media is a prime example of why social media companies need to moderate and mitigate the spread of misinformation on their platforms.

Broadly, instances of misinformation related to the COVID-19 pandemic led to the spread of various harmful practices, such as facilitating stigma and apprehension surrounding the vaccine and creating issues for vaccination efforts by many nations. Prominent misinformation stories circulating online included that vaccine trial participants had died after taking a candidate COVID-19 vaccine and that the pandemic was a conspiracy or bioweapon (Geldsetzer, 2020). A study published by *Nature* found that misinformation influenced people's behavior during the COVID-19 pandemic in the U.S. and UK in three significant ways: it made people less likely to report willingness to get vaccinated against COVID-19, it made them less likely to recommend vaccination to vulnerable people in their social circle, and it decreased people's willingness to comply with public health guidance measures (Loomba et al., 2021). This is yet another example showing how the mediation of disputable content by social media platforms is critical, and how the lack thereof can lead to real-world consequences. The moderation of health misinformation is critical for ensuring verified information is disseminated to social media users, as well as protecting public safety and health.

Hate Speech

When platforms remove specific posts or an individual's access to their accounts, they risk being accused of limiting the speech of their users. These decisions are made even more complex when determining whether content qualifies as hate speech. Adam Klein (2017) refers to hate speech online as something that "can be defined as the strategic employment of words, images, and symbols, as well as links, downloads, news threads, memes, conspiracy theories, political blogs, and even pop culture, all of which have become the complex machinery of effective inflammatory rhetoric." Hate speech is not always easy to detect under this definition. For instance, if a user created a post that included contentious rhetoric, they could argue that their post was a joke not intended to express hateful or harmful speech and, therefore, should not be taken down. Circumstances such as these might allow a post with inflammatory content to remain visible to the public because whether it can be categorized as hate speech is disputable.

This dilemma, described by Klein as the "Spectrum of Debate" and "Spectrum of Hate" question is a primary barrier for the prevention of hate speech online (Klein, 2017). This spectrum asks: *How does one determine what is hateful speech and what is strongly worded political or cultural speech?* Even if a content moderator might suspect that something has undertones of hate speech, the very structure of social media lends itself to prioritizing keeping content visible if that content is disputable (Jakubowicz, 2017).

For example, Media Matters discovered in 2022 an Instagram account whose primary goal is to post about local violent crimes. However, this account was dedicated specifically to posting about crimes allegedly committed by Black people against White people, a purpose which has clearly racist underpinnings. That being said, the account did not violate any of Instagram's community guidelines in the text of its posts and has therefore been allowed to remain active (Carter, 2022). This example demonstrates the way in which social media platforms will allow inflammatory and racist content to remain visible if that content is disputable. The reason for this lies in the ways in which social media companies were founded.

Many of the largest social media platforms such as Facebook, Instagram, and YouTube maintain commitment to democratic values of free expression and speech within their platforms, which can allow for disputable content to exist on their platforms (Jakubowicz, 2017). This ideology is due in large part to these corporate entities being formed in the United States. While these companies are also driven by

revenue and the desire to avoid legal litigation, they will typically prioritize free expression if it does not compromise their profit. The challenge therefore is that if a post cannot be definitively labeled as hate speech, then platforms are inclined to keep the post visible to maintain the free expression they value.

Extremism

Moderating violent extremist groups on social media presents a major problem in both the vast amount of such content and issues regarding freedom of expression. Extremist groups use social media in order to further their political goals. However, arguments regarding free expression on social media and detection of extremist content make moderating the activities of these groups difficult. The lack of proper moderation can lead to violence in the real world against both individuals and groups.

Hate speech and recruitment into extremism are two types of disputable content that can lead to much larger problems for society, namely violence and political division. What begins as hate speech can quickly become a call to violence, and those recruited into extremist groups are likely to answer these calls. Incitements of violence on the internet are both heavily moderated by content moderators and outlawed in most democracies. However, disputed content that includes incitements to violence can be difficult to detect.

In her piece on how individual actors use social media to organize political action and criminal action, Kaminski (2013) gives the example of a Twitter user who tweets about a gathering of individuals intending to rob a bank. Should this robbery occur, the tweet could be used as direct evidence that a crime has been committed (Kaminski, 2013). However, should the robbery not occur one might argue that the tweet did not incite violence, it only created the potential for violence. In this sense content moderators must ask themselves if a disputed post calls for violent action or simply uses language to exaggerate or embellish a point.

When extremist organizations post disputable content that includes language targeting individuals or groups, platforms are forced to decide whether that language calls for violence or not. Failure to effectively moderate disputed content could lead to the destruction of property, injury to individuals or groups, and even death in extreme cases. However, if the disputed content is removed there is also the potential that a user's speech was censored. Balancing effective moderation and censorship is a prime example of the dilemma social media platforms face when dealing with the content posted by White supremacist and terrorist organizations.

Both types of organizations, which exist in a variety of geopolitical contexts, use social media for their own ends and post disputable content. While some see their content as illegal or wrong, the organizations and their supporters view it as justifiable. White supremacist and terrorist content will be used as examples to better

understand how extremist groups create disputable content, particularly regarding hate speech and recruitment into extremism.

Disputable Content Dissemination: Extremist Groups

Terrorist organizations' use of social media to recruit new members is a concern for nation-state actors (Hossain, 2015). Many people are familiar with the explicit content posted by terrorist organizations such as executions; however, fewer people understand the ways in which terrorist organizations use social media to complete goals such as recruitment. While the recruitment activity of terrorist organizations online may not be very visible to the general public, terrorist organizations benefit from using social media to train, recruit and communicate with members (GCTC, 2021). Part of the reason recruitment is mostly unnoticed by the general public is that the most extreme terrorist group activity occurs on encrypted private messaging platforms, such as Facebook Messenger (Wray, 2017). Recruiters for terrorist organizations will post content on platforms such as Facebook or Twitter knowing that it will be eventually taken down in order to discover individuals who may be engaging with the content, such as watching a video, liking, or commenting on a post (GCTC, 2021). Once a recruiter has found a user engaging with their visible content they will invite the individual to an online encrypted messenger forum, like Facebook Messenger, in order to attempt to begin a conversation. These efforts are targeted at specific individuals or groups. One 2018 study by RAND Europe found that students at the University Medical Sciences and Technology in Khartoum, Sudan were being targeted for recruitment efforts by extremists from ISIS (Ward, 2018). These targeted interactions demonstrate how social media directly enables terrorist organizations to attempt to recruit and further their organizations' goals (Speckard, 2020).

Under pressure from governments, large online companies like Facebook and Twitter are attempting to severely limit the threat of terrorist activity on their platforms (Macdonald, 2019). However, in their efforts to curb terrorist activity these social media companies have instituted policies which have negatively impacted users. Policies created to limit ISIS propaganda disproportionately impact the speech of Muslim and Arabic-speaking users (Díaz, 2021). Human content moderators are subject to implicit bias like everyone else, which is "a form of bias that occurs automatically and unintentionally, that nevertheless affects judgments, decisions, and behaviors" (National Institutes of Health, 2022). Additionally, AI content moderators, programmed by human beings, are also influenced by implicit bias. Both human and AI content

moderators struggle to differentiate between ordinary religious Islamic practices and language and more extremist ideologies because such moderators were trained by companies who have not worked to understand the difference between the two. A content moderator who does not have experience differentiating between religious practice and extremism might flag a user's speech even although it is not harmful (Torok, 2021).

The bias in moderating content creates issues related to censorship and freedom of expression within social media platforms as Muslims are forced to self-monitor their social media activity or risk being labeled as engaging in extremist behavior by content moderators. Such a case occurred on TikTok in 2019 to 17-year old Feroza Aziz (Ma, 2019). After posting videos condemning China's treatment of Uyghur Muslims, Aziz's account was suspended. However, TikTok claims the suspension was due to her posting of a satirical video making fun of Islamophobic comments she had received as a Muslim-American. The video Aziz posted included a photo of Osama Bin-Laden for satirical purposes, however TikTok claimed this video had to be taken down for violating its "anti-terrorism" policies. For Aziz, a policy meant to keep extremist content off the TikTok's platform directly impacted her free expression. In this sense, social media platforms must grapple with how to address the use of social media by extremist organizations while giving room for ordinary people to engage freely and without fear of bias on social media platforms.

White supremacist and far-right groups also use social media platforms to recruit into their organizations. The Anti-Defamation League (ADL) argues that these extremists seek to recruit younger college-educated white individuals who may be vulnerable to indoctrination. ("Alt Right", 2016). The recruitment of this demographic most frequently occurs through the "alt-right pipeline", which is defined by the Berkeley Political Review as "the path people follow as they engage with online content, forums, and peers that take them further to the right-wing political extreme" (Mariash, 2022). While still intentional on the part of these organizations, recruitment into the extreme right and White supremacy groups through use of this alt-right pipeline is typically less direct than the recruitment of terrorist organizations. This recruitment slowly brings individuals into extremism using memes and humor ("Alt Right", 2016). The method of memes, defined as "an image with caption text" (Wiggins & Bowers, 2015), creates an environment in which content moderators cannot effectively moderate extremist content because it is hidden under the guise of humor.

Furthermore, a person might view a meme posted by a White supremacist account and be taken to more extreme content through social media's use of

algorithmic content sharing. Algorithmic content sharing is social media platform's use of algorithms to share new content to users based on that user's past behavior (GIFCT, 2021). On all platforms these algorithms typically collect information based on the user's previous behavior to recommend new content that will be most engaging (GIFCT, 2021). Therefore, when a person engages with content related to the alt-right, they will likely be recommended more content related to the alt-right.

One of the memes utilized by White supremacists online for recruitment through algorithmic content sharing has been the "Pepe the Frog" meme. Originating as an American pop culture image, Pepe the Frog is not inherently racist. However, social media users noticed an uptick in White supremacist related Pepe images. This increase was intentional on the part of White supremacists, who purposefully began creating Pepe images with White supremacist messages in them (Domonoske, 2016) (see Figure 2).

Figure 2: Pepe the Frog Images

For this reason, a person who engages with a Pepe meme created by a member of a White supremacist group might then be suggested more content from extremists even if that is not content that they had wanted to engage with. This relates to issues of content moderation as there was nothing inherently wrong with the image of the Pepe frog alone when it was originally created, so content moderators were unlikely to flag this content as harmful even though it could eventually lead individuals to extremism. Today, Pepe the Frog's use by White supremacists is known and social media platforms have taken steps to address its use (Farokhmanesh, 2018). However, in the future other images or culture references could come to be used by White supremacist organizations for similar purposes. Moving forward, social media

companies must grapple with balancing free expression and disrupting extremism on their platforms.

To conclude, both terrorist organizations and White supremacists use social media platforms in order to recruit into their organizations. Issues such as implicit bias and free expression further complicate efforts to disrupt recruitment on social media platforms. Content moderators then must attempt to disrupt these recruitment efforts while balancing the nuance of meme and internet culture while avoiding censoring regular users.

Private Sector Behavior: Common Moderation Methods

As social media platforms like Meta's Facebook and Instagram, Twitter, YouTube, and TikTok continue to grow and attract millions of users, the issue of content moderation has become increasingly important. These platforms use various techniques to monitor and moderate content, including automated algorithms, human moderators, and user reporting systems. Despite their best efforts, however, these platforms often struggle to strike the right balance between allowing free expression and preventing harmful content such as hate speech, fake news, and violence. Recently, controversies have arisen over how these platforms have handled moderation, highlighting the need for continued discussion and reform.

Despite the guidelines and efforts made by major platforms in terms of content moderation, these companies face concerns surrounding users' values regarding freedom of expression and usage on these platforms. Patterns of missed and unregulated misinformation and disinformation, hate speech, and extremist content on these platforms demonstrate a lack of transparency when it comes to moderating what is considered disputable content. Additionally, accountability measurements that these platforms put on users when it comes to users breaking community guidelines also create complicated discussions around addressing repetitive behaviors of violating conduct over and over.

Although there are AI detectors, human moderators, individual users' reports, and other preventive measures to combat forbidden content on the platforms, the recurrences of disputable content through various forms of media can spiral from one post to another. The five social media platforms of focus experience over one million estimated daily posts, which range from benign posts to areas of disputed content. Hence, moderating the nuances and subjectivity of content on the platform is challenging, especially regarding areas of content that are difficult to both identify and categorize within community guideline criteria (e.g., violent content).

Here we discuss moderation tactics that are shared across Facebook, Instagram, Twitter, YouTube, and TikTok. Through our research, we identify the most similar processes as preventative measures, AI and human moderators, and issues with biases throughout moderation approaches.

Preventative Measures: Age Verification

The most common preventative measure for content moderation across major social media platforms is age verification when creating accounts (Carson, 2022). Users

must enter their date of birth during the registration or sign-up process and agree to abide by the platforms' terms of service. All five platforms require users to be at least thirteen years of age to utilize them, which is primarily due to the U.S.'s Children's Online Privacy Protection Act (COPPA) (Nguyen, 2019). Many countries have adopted this age restriction because of COPPA, which aims to protect children's privacy online.

In addition, to comply with COPPA, YouTube requires a user to provide information such as a government-issued ID, passport, or credit card to verify that they are over 18 and can view age-restricted content (YouTube, n.d.). TikTok also has implemented age restrictions on certain types of content, which means that users who are under 18 may not be able to access specific videos or features (TikTok, n.d.). Since 2021, TikTok has made underaged accounts private by default, meaning that only people verified by the account holder can follow or access their content (Perrie, 2022). Twitter also requires users to verify their age before following other accounts if they still need to provide it when signing up for the platform (Twitter, n.d.). Meta plans to adopt more AI technology and increase collaboration with operating system providers and internet browsers to strengthen its age verification process (Finkle, 2022).

Despite these big platforms' efforts to protect minors from harmful content, limitations also exist in these platforms' age verification processes. Users may need to provide more accurate information about their age, making verifying accuracy difficult for the platforms. Given that it is not illegal to lie about their age on social media platforms, a lot of the underage population can quickly enter the platforms anyway. For instance, YouTube only makes someone verify their age if they use the platform in certain regions such as Australia, the European Union, European Economic Area, Switzerland, or the United Kingdom (Google, n.d.). If a user is not in any of those areas, they can lie about their age without ramifications, making it more challenging to get around the verification. Some users may even create fake profiles or use a parent's account to get around age restrictions (Sundaravelu, 2022). The existence of loopholes that enable the circumvention of age verification measures implies that such measures do not effectively safeguard all minors from accessing age-restricted content. While laws like COPPA and similar regulations in other countries are in place to protect children, it is still significant for social media platforms to continuously improve their age verification processes and take steps to ensure that minors are protected from harmful content online.

Preventative Measures: User Ability to Flag or Report

All five platforms allow users to anonymously flag or report other users' content. Anonymous reporting can be a valuable tool for people who want to bring attention to problems or to violate content without fear of retaliation. With millions of users, social media platforms can effectively moderate a large amount of content by enabling users to report any content they deem inappropriate or violating the platform's community guidelines. By leveraging user reports, social media can expand their reach in moderating a more comprehensive range of content that may require attention. Also, this system can give users a voice in helping to shape the content they see on the platform.

However, when reporting is anonymous, it can be difficult for the platform to assess a report's validity and take appropriate actions. This can lead to inconsistent enforcement of the platform's policies and undermine the effectiveness of the reporting system. Anonymous reporting can encourage cyberbullying or trolling, as people may feel empowered to make scathing reports without facing the consequences (Castella & Brown, 2011). This can discourage positive engagement on the platforms. In summary, while anonymous reporting can be a valuable tool for moderating content on social media platforms, its limitations and potential for abuse must be carefully considered to ensure a safe and positive online community for all users.

Preventative Measures: Warning Labels

Warning labels on social media platforms come in different types and serve different purposes. One type of warning label aims to protect users from viewing potentially distressing content, such as graphic images of violence or death. Warning signs indicate potentially graphic, inappropriate, or sensitive materials, allowing users to decide whether to view the content (Emebo, 2019). All five platforms we analyzed also utilize warning labels to provide additional context or information about content that may not be suitable for all viewers. For example, users on Twitter are required to click through a warning notice to view a tweet that contains sensitive information, and such tweets will not be recommended across the service (Roth, 2022). A similar precaution was taken by TikTok, as they implemented warning labels in 2021 to alert viewers and remind users who attempt to share clips that a particular video has been flagged as unverified content and notify users when a warning label has been added to their clip (Hutchinson, 2021).

Another warning label informs users that the information presented may be inaccurate or misleading. It helps users make more informed decisions about the content they view and believe on social media platforms. However, Facebook failed to add warning labels on 180 million pieces of content before the 2020 U.S. Presidential Election (Kraus, 2020). Despite warnings from election officials and fact-checkers, false claims about widespread voter fraud and claims of premature declaration of victory were still widely shared on the platform. This demonstrates that simply adding warning labels is not enough to combat the spread of misinformation effectively.

Additionally, warning labels can have a limited impact, as users may ignore them or be more likely to engage with the flagged content. Labeling is only effective if warning labels are enforced, and users take them seriously. There is often a lack of enforcement mechanisms to ensure that users abide by the labels, and some users may ignore the warning notices and view flagged content anyway. The purpose of warning labels is to give users a choice to see or avoid content. However, the concern is that some users may ignore the warning labels and still share flagged content, which could lead to the continued spread of misinformation or harmful content. While the labels give users more control over what they see on the platform, their effectiveness also depends on users' willingness to comply with them. In addition, once a piece of false information has spread widely across social media platforms, it can be very challenging to retract or counter it, even by implementing a warning label. In conclusion, the limitations and inconsistencies of these preventative measures highlight the need for a multifaceted approach that integrates technical solutions, human moderation, and user education.

AI and Human Moderators

Understanding how AI and human moderation work is integral to the broader discussion of social media platform's role in the conversation around internet freedom in democracies. This section will discuss an overview of commonalities in Facebook, Instagram, Twitter, YouTube, and TikTok's content moderation process and describe how content is flagged for moderation. Specific characteristics of how each platform runs its content moderation process will also be discussed to understand the complexities within AI and human moderation.

Overview of the Moderation Process

Most popular social media platforms—including Facebook, Instagram, Twitter, YouTube, and TikTok—rely on three effective methods to help moderate content and users on their sites: automated moderation software, human content moderators, and its users (Congress, n.d.). Simply put, most prominent social media companies' human and automated moderation systems adhere to the following formula: artificial intelligence automatically reviews every piece of content posted on the platform, while the human moderators review content that has either been flagged by the AI tools or reported by a user of the platform (Tarasov, 2021). The human moderators are hired directly by the company or through a third-party contractor, or a mixture of both. For example, Facebook and Instagram's parent company, Meta, hire moderators directly and through third-party companies like Accenture and Voxpro (Shead, 2020). TikTok only hires moderators directly for their company (Shead, 2020).

The relationship between the human moderators and AI technology is often viewed as "collaborative" since the AI technology continues to learn and modify how it detects content that violates platform policies based on the human moderators' feedback regarding flagged content (Halprin, 2022). However, the downside to this "collaborative" relationship is that there seem to be no safeguards in place to ensure the AI moderation technology does not learn biases through its human counterparts. Another downside to this relationship is the emotional and psychological toll human moderators face when they moderate thousands of pieces of content that contain harmful images, videos, and rhetoric missed by the AI systems (Arsht & Etcovitch, 2018).

It is difficult to find information regarding social media platforms' moderation systems that do not come directly from the companies. The reality of how social media companies moderate the content on their respective platforms may work differently than the available information suggests.

Identification of Moderated Content

The identification of prohibited content on social media platforms is convoluted and complex. Automated moderation across all social media platforms is not always accurate. There have been documented instances where these systems flagged materials that did not violate content policies. One such instance occurred during the beginning of the COVID-19 pandemic when YouTube sent their human moderator employees home to stop the spread of the virus. Because of this, YouTube notified its

content creators that their content might be removed or flagged more often for violating community guidelines as the company had to pivot and rely more on automated moderation, which, as stated, is not always accurate (Leskin, 2020).

Conversely, moderation systems have failed to remove all content that violates community guidelines. Facebook's AI content moderators had trouble detecting posts containing COVID-19 conspiracies, which led to violative posts remaining visible for all users (Patel & Hecht-Felella, 2021). The AI moderation tool's human counterparts are not error-free either. The final decision to remove a piece of content is often up to the discretion of the companies' human moderators, though they must follow their specific content policies in place (Stackpole, 2022). These policies are often subjective and disputable. Individual companies do not readily share or release detailed information about how their human moderators are trained to operate in the context of disputable content.

For example, like other social media platforms YouTube trains its human moderators to differentiate between violative and non-violative material, although YouTube does not clearly state what is included in the training (Halprin, 2022). YouTube has, however, given small glimpses into how its human moderators are taught to review flagged content through previous content moderation scandals. Take the 2019 Crowder and Maza incident, in which Steven Crowder, a conservative political commentator, posted videos on his YouTube channel, calling Maza an "angry little queer", "gay Mexican", and "gay Latino from Vox" (Koebler & Lamoureux, 2019). Despite acknowledging that Crowder's comments were hurtful, YouTube stated that they would not remove or penalize Crowder's channel, regardless of public pressure, because his videos did not violate their content policies (Koebler & Lamoureux, 2019). When human moderators review a grey content area, YouTube explains that they consider the "context of the entire video" (Koebler & Lamoureux, 2019). In the Crowder and Maza hate speech example, YouTube further stated that they considered whether the content in question leaned more towards debating public opinions or was only published with malicious intent (Koebler & Lamoureux, 2019). They concluded their statement by explaining that they apply these policies regardless of the number of views a video has, thus implying that YouTube content moderation is not biased towards popular videos or creators (Koebler & Lamoureux, 2019). Although it is difficult to understand the full scope of how YouTube moderates content with little available information, instances like Crowder and Maza illustrate that YouTube is aware of the intricacies of content moderation and is taking small steps to address it. However, it would be more effective for YouTube to be transparent about its moderation training and techniques without inciting public outrage beforehand.

Biases and Inequality

As previously established, most platforms rely heavily on automation as their main form of moderation, while some also employ teams of human moderators to analyze flagged content and monitor the platform as a whole. In terms of the technology aspect of moderation and general engagement with social media, platform algorithms. or the personalized content suggestions, are strengthened by user participation with suggested content from artificial intelligence. Disputable content like hate speech, misinformation, and propaganda can be detected using NLP, or natural language processing concepts, to analyze sentences' lexical (of or relating to the words or the vocabulary of a language) and syntactic (relating or according to the rules of syntax) features (Khanday, 2022). Also, as aforementioned, users play a role in the moderation of content with the ability to flag or report content that may be harmful. However, the human inclination to interact with content that can be categorized as sensationalist and outrageous can often result in moderation systems learning to prioritize disputable content (Diaz, 2021).

Modern AI processes can learn racial, gender-based, and linguistic biases as they analyze user engagement with certain content, making them fallible to disputable content, which tends to generate the most engagement because of the sensationalism of the content (Diaz, 2021). Understanding that "bias is a reflection of the data algorithm authors choose to use" indicates why certain vulgar posts which violate community guidelines, such as those made by Kanye West or White supremacist-related groups, go unnoticed by both the technology and team of human content moderators (Nelson, 2019). It was also revealed in 2019 that terrorist content was slipping through YouTube's AI moderation systems because Arabic was used in the content (Ayad, 2019). As the failures of the current system continue to rise in public discussion, users find more and more ways to post disputable content while curbing the potential consequences of posts being flagged or taken down, and accounts being suspended or banned by the platform.

Language and Personal Bias

Alternative language is when users will create alternative words which sound like and represent harmful language but will not be flagged by the software as a violation of guidelines (Tait, 2022). For example, users on TikTok will post a video discussing topics relating to death but will use the word "unalive" in place of the word "kill" so that the video does not get taken down by the platform (Tait, 2022). Once

those words are learned by the intelligence software, they will be taken down accordingly, however, users are constantly updating their use of alternative language to avoid the consequences of violating platform guidelines.

The use of alternative language is a particular issue when it comes to hate speech on the internet, as that alternative language is not flagged, yet it is still widely understood as hate speech by users of the platform. Additionally, oftentimes when users with high social influence and a large following post disputable content or content which violates platform guidelines, the platforms will bend the rules to accommodate those users (Baker-White, 2022). After all, the high-profile users are the ones generating profit and engagement for the companies, which proves to be a weak spot in maintaining equal standards for users in relation to content moderation.

YouTube

YouTube is home to millions of hours of video content and is one of the world's biggest platforms, with nearly 2.5 billion users logging on to use the platform each month (Statista, 2022). The popular video-sharing site was created in 2005 by three former PayPal employees who received funding from venture capitalists for their technology startup, which would later be named YouTube (Helft & Richtel, 2006). YouTube is available in 190 countries, but only 91 have local versions (Map of YouTube Availability, 2022). Local versions of YouTube allow users to view country-specific content, a fully translated website in the respective language, and monetize their content (Brouwer, 2015). As of August 2022, YouTube has been banned in five countries: China (excluding Hong Kong and Macau), Iran, North Korea, Turkmenistan, and Eritrea—which has experienced intermittent banning since 2011 (BBC, 2012) (Liebelson, 2014) (Ferghana.ru, 2009).

YouTube was the first video-sharing specific social media platform. In 2006, YouTube was bought by Google for $1.65 billion (Heft, 2006). Under Google's ownership, YouTube has undergone an immense transformation and growth, mainly through its implementation of ads and premium subscriptions (Downey, 2021).

Community Guidelines

Like other popular social media platforms, YouTube has established a set of community guidelines that outline the rules users must follow to use the platform. YouTube's Community Guidelines policies fall under five umbrella categories: spam and deceptive practices, sensitive content, violent and dangerous content, regulated goods, and misinformation (YouTube, n.d.). Under each of the five categories are sub-categories that delve deeper into what content is and is not allowed on the platform. With its community guidelines, YouTube aims to “create a safe community where its content creators can freely express their creativity, experiences, and perspectives” (YouTube, n.d.).

While YouTube has publicly declared its community guidelines and made them readily accessible on its website, the information it displays publicly can be misleading and is often not representative of the whole picture. YouTube's Community Guidelines should be viewed as the projection of the polished image that YouTube wants the public to see, rather than as the whole truth.

YouTube has attempted to create a safer environment for its users by implementing a "three-strike" rule regarding users' channels and content moderation.

In short, if a channel receives three strikes, the channel may be deleted or deactivated. If a user is revealed to have violated YouTube's Community Guidelines, they will receive either a warning for first-time offenders or a strike against their channel (Google, n.d.). As stated before, the user can have a human review of the content deemed violative against YouTube's policy. However, if the human moderator believes the content is prohibited on YouTube, the channel will receive a strike (Google, n.d.). For each "strike" a channel receives, specific YouTube "privileges" will be taken away for a predetermined time. For example, if a channel receives its first strike, it cannot upload videos, live streams, or stories for one week (Google, n.d.). The consequences become steeper and more prolonged the more strikes a channel receives.

Another strategy YouTube has implemented to hold users accountable is demonetizing content that violates their community guidelines. While this measure does incentivize users to follow content policies since many times the creators' livelihoods are attached to how much money they make from their YouTube channel, the issue with demonetization lies in that not every YouTube channel is eligible for monetization, also known as YouTube's Partner Program (YPP) (Google, n.d.). YouTube has specific quantitative measures that a channel must reach before it can join the program. YouTube requires a channel to have 1,000 subscribers and either 4,000 valid public watch hours in the last 12 months or 10 million valid public short-form content views in the last 90 days to be eligible for YPP (Google, n.d.). Additionally, only 122 out of 190 countries where YouTube is available are eligible to participate in YPP (Google, n.d.). This means channels that have not yet met the necessary subscriber or view threshold and users in the 68 countries without YPP will not face the consequences of demonetization. Further, the consequences for channels that are not monetized and post disputable content are less severe than those that are monetized and post disputable content.

YouTube moderates and reviews all content posted on the platform, including videos, comment sections, thumbnails, and external links (YouTube, n.d.). YouTube has traditionally removed content that violates its community guidelines or other content policies. On its community guidelines page, the company states that it applies its content reviewing and moderating services to every user equally, regardless "of the subject or the creator's background, political viewpoint, position, or affiliation," however that has not always been the case, as will be explored in later sections of this report (YouTube, n.d.). While the type of content moderated on YouTube might seem straightforward (i.e., content that violates YouTube's policies will be removed), it is far from it. Like many social media platforms, YouTube has struggled to find the

most efficient and straightforward way to define what type of content is moderated due to the disputable nature of the content and vague policies. YouTube also struggles to outline appropriate and effective exceptions to its content moderation policies. YouTube's struggle is not unique and is part of the greater struggle surrounding internet freedom of expression affecting all popular social media platforms.

Disputable Content: Disinformation and Misinformation

According to YouTube's Community Guidelines, misinformation is categorized as misleading content with a severe risk of real-world harm (Google, n.d.). However, YouTube's attempts to curb this issue are not always successful. YouTube uses examples such as the promotion of harmful remedies or treatments, manipulated content, or content interfering with democratic processes as misinformation content not allowed on the platform (Google, n.d.). YouTube further breaks down four sub-categories within this section of its community guidelines: misinformation policies, election misinformation policies, COVID-19 medical misinformation policies, and vaccine misinformation policies (Google, n.d.).

One of the most relevant and current incidences that called YouTube's misinformation policies into question was the outbreak of COVID-19. Li found that one in four spoken English videos on YouTube regarding COVID-19 contained misinformation (Li et al., 2020). Researchers began reviewing YouTube content regarding COVID-19 in March 2020. They narrowed down the videos based on factors such as only looking at spoken English videos or discounting duplicate videos (Li et al., 2020). In the end, researchers looked at 69 videos— of which the total views added up to 257, 804,146 —and found that only 50 included factual information. At the same time, the other 19 contained misleading or inaccurate information about COVID-19 (Li et al., 2020). The 19 videos containing misinformation received 62,042,609 views, or about 24% of the 69 videos' views combined (Li et al., 2020). Researchers also found that government agency and professional videos on COVID-19 were the most factual and accurate but received significantly fewer views than other sources (Li et al., 2020). They found that the videos with COVID-19 misinformation mainly came from entertainment, network and internet news sources, and consumer videos (Li et al., 2020).

However, contrary to Li's findings, 2021 YouTube claimed to have made a significant effort to remove COVID-19 misinformation from the platform by removing one million videos containing misinformation on COVID-19 since February 2020 (France-

Presse, 2021). While Google's statistics may be accurate, the research report shows that misinformation is still finding its way past YouTube's moderation systems. Further, it is crucial to understand that the harmful ideas in unmoderated misinformation content have already reached a broad audience in the time between its upload and removal, which spread off the platform and manifest into more catastrophic consequences like influencing people using harmful medical advice (Li et al., 2020).

Disputable Content: Hate Speech

YouTube's policies on hate speech are often the center of the debate over freedom of expression on the internet. YouTube defines content regarding hate speech and violence as content promoting violence or hatred against individuals or groups based on specific attributes like gender, race, or religion (Google, n.d.). The company also includes content that contains predatory behavior, graphic violence, and malicious attacks under this policy (Google, n.d.). However, due to the often-subjective nature of this category, YouTube has faced controversies over its hate speech and violent content moderation choices (De Vynck, 2021). YouTube does not include any information about freedom of expression in its community guidelines or content moderation policies. A frequent critique of YouTube's content moderation, in general, is that it violates freedom of expression—a key component to any functioning democracy. That said, YouTube is part of a private company that functions on its own rules and values, which are not required to align with democratic values. YouTube is protected from having to incorporate freedom of speech in its content policies. However, many argue that freedom of speech is something YouTube must consider when creating content moderation policies (BBC, 2020).

In 2019, YouTube faced backlash over a creator's hateful and offensive comments about a Vox journalist, Carlos Maza. As discussed previously, a conservative political commentator with 3.8 million subscribers named Steven Crowder posted videos on his YouTube channel calling Maza an "angry little queer", "gay Mexican", and "gay Latino from Vox" (Koebler & Lamoureux, 2019). Although YouTube stated that Crowder's comments were hurtful, the company did remove the videos or penalize Crowder's channel, stating that his videos did not violate their content policies (Koebler & Lamoureux, 2019). After a further review of Crowder's channel conducted after public pressure, it was eventually decided that rather than remove Crowder's channel or the videos in question, Crowder's channel would be demonetized (Koelber & Lamoureux, 2019). In other words, he would no longer make

money through advertisements on his videos. YouTube explained its decision by claiming it did not find evidence that he violated its hate speech policies since he did not "urge his followers to go after Maza" in any of his videos (Koebler & Lamoureux, 2019). However, as explained before, YouTube's policies state that they will remove content that promotes hatred against an individual's sexuality, race, or ethnicity (Google, n.d.). It is unclear why Crowder's comments did not fall under this rule since, in the context of his videos, Crowder was using Maza's identity as a slur. This scandal is partly due to YouTube's ambiguous terms and language around its hate speech policies. If it was clearer what content qualifies as hate speech on YouTube, it is possible that a more effective and publicly satisfying solution would have been reached.

Two reporters who covered this specific YouTube scandal, Jason Koebler and Mack Lamoureux, stated that YouTube's approach to moderating hate speech was that "YouTube has decided that Crowder's content is too hateful for advertisers, but not too hateful for the platform itself" (Koebler & Lamoureux, 2019). Further, YouTube's handling of this situation brings into question how impactful the demonetization of Crowder's channel was compared to removing Crowder from the platform overall. From this situation, it is possible that YouTube was more concerned about its advertisers and financial gain than protecting the well-being of its users. It should also be noted that YouTube only addressed this situation once they faced public pressure, thus begging the question if Crowder's channel would ever face repercussions without public scrutiny and how many situations like this occur that are not addressed by YouTube.

Disputable Content: Extremism

Extremism falls under YouTube's violent or dangerous content policy as a sub-category (YouTube, n.d.). It specifies that content that promotes, aids, or applauds violent extremist or criminal organizations is not allowed on the platform (Google, n.d.). While YouTube has improved its policies regarding extremist content on YouTube, they recently partnered with the Anti-Defamation League to remove such content, it is by no means a perfectly effective system for filtering out all extremist content (ADL, 2016). Loopholes in which extremist content continues to get uploaded to the site exist. YouTube's first line of defense, its AI moderation technology, does not detect every piece of extremist content, a common issue across the platform. The mixture of biases, gaps, and content volume complicates managing this issue, while the consequences of extremist content on YouTube are high due to

radicalization and real-world violence. A notable example is YouTube's role in radicalizing the Christchurch perpetrator, who claimed that extremist content on YouTube played more of a role in building his extremist ideas than extremist far-right sites (D'Anastasio, 2020).

YouTube allows content that would typically violate their rules but is used in an educational context to be uploaded onto the platform. Extremist groups have become aware of this loophole and have taken advantage of it. For example, Jihadists have learned to tag or title content as education or include phrases related to education, which often allows the content to bypass AI moderation (Ayad, 2019). They have also learned to link their extremist content to mainstream content on YouTube. For example, an ISIS or Al-Qaeda member may make a video playlist on YouTube and mix in extremist content with more mainstream content like popular Arab music videos or belly dancing, thus encouraging more views of the violent content (Ayad, 2019). Although this example looks at extremist groups from the Middle East, the use of YouTube as a platform to spread extremist ideas is not exclusive to groups like ISIS or Al-Qaeda. Many other extremist groups, like White supremacists or neo-Nazis, use YouTube to spread their ideology as well (Greer & Ramalingam, 2020).

A team researching terrorist content on social media platforms found that searching in Arabic for the names of terrorist ideologues, groups, and texts showed users extremist, violent content (Ayad, 2019). YouTube's AI technology and human moderators could not detect and remove the content. In 2019, YouTube stated it removed 89,968 violent videos out of one million suspected of containing extremist content within the first three months of the year (Ayad, 2019). It is unclear whether the 89,968 videos represented a process that suggests only the most extreme videos were removed or, instead, the selection process was over-generalized. While this is a positive step in the right direction, it begs the question of how much content still needs to be noticed by YouTube's moderation systems because of factors like language, misleading tagging or titling, or other intricacies. These issues are fundamental for YouTube to solve because of the high possibility of real-world harm associated with extremist groups.

In more recent efforts to combat extremist content, with the help of Moonshot, a tech startup, and Google's Jigsaw, YouTube implemented a pilot project called the "redirect method" (Greer & Ramalingam, 2020). The redirect method uses advertisements on Google to reroute users viewing extremist content to videos and content that counter those extremist narratives (Greer & Ramalingam, 2020). The method was used on content regarding terrorist groups in the Middle East and White supremacist groups, and was considered successful (Greer & Ramalingam, 2020).

During the pilot run, users searching for White supremacist content consumed 5,509 minutes of videos countering White supremacist ideology (Greer & Ramalingam, 2020). Users searching Islamist-inspired extremist content viewed 534 minutes of counter-narrative content (Greer & Ramalingam, 2020). Although the pilot program found the redirect method effective in combating extremist content on YouTube, it has not yet been tested outside the U.S., nor been implemented permanently (Greer & Ramalingam, 2020). It is impossible to conclude if the positive results would be replicated in other regions or continue to be effective with long-term use. Additionally, the researchers overseeing the redirect project could not conclude if the counter content helped push the users away from extremism or if users were "hate-watching" the counter-narrative content (Greer & Ramalingam, 2020).

TikTok

Developed and owned by the Chinese Bytedance, TikTok is a video-sharing app alternative to social media sibling to the popular Chinese-developed app Musical.ly (Doyle 2023). The platform has accumulated 100 million users in the United States alone since its debut there in August 2018, although globally, TikTok has around one billion users (O'Connor, 2021) (Doyle, 2023). Today, TikTok is one of the most popular social media platforms, with 850 million downloads, 689 million monthly active users worldwide, and over one billion video views per day (Kang, 2022). As the most downloaded app of 2020 and available in over 154 countries and 75 languages, the platform is only continuing to grow in popularity (Lin, 2021).

Despite its quickly rising in popularity, many countries are suspicious of the TikTok China-based company, ByteDance's relations with the Chinese State (Espada, 2023) (Milmo, 2022). This suspicion led to bans of the app on government devices in countries like the United States, Canada, and Taiwan (Berman, 2023). Other countries such as Afghanistan, Armenia, Bangladesh, India, Indonesia, and Iran have banned the app across the country due to TikTok's potential to spread inappropriate content (Berman, 2023). With the United States having the majority of TikTok users, there have been rising geopolitical and economic concerns between the U.S. and China over leadership, data security, and national security (Kangs, 2023).

Community Guidelines

A platform like TikTok sees millions of posts daily and regulating all that content on the platform can be challenging. The company has made numerous announcements on improving the platform's moderation system, including greater transparency around their moderation so that users are able to experience a creative and safe environment (Walker, 2023). The promise to promote TikTok's creative values of "[inspiring] creativity and [bringing] joy" can be found in the company's mission statement (TikTok, n.d.). To reflect these values, TikTok's guidelines address 13 major sections, which include: minor safety, dangerous acts & challenges, suicide, self-harm, disordered eating, adult nudity & sexual activities, bullying & harassment, hateful behavior, violent extremism, integrity & authenticity, illegal activities & regulated goods, violent & graphic content, copyright & trademark infringement, platform security, and ineligibility for the "For You Page" (FYP) (TikTok, n.d.). TikTok has publicly announced that the platform will continue to improve and expand its moderation policies to prevent content, like disputable content, from causing real life

implications (Keenan, 2022). Despite these many categories, content moderation posed by TikTok needs more transparency regarding the kind of regulated posts taken down.

Social media platforms like TikTok can function as an echo chamber for user information, which continuously and progressively depends on user-generated content and recommendation algorithms to generate personal For You Pages (FYP) for its users. Based on an article on "HER CAMPUS", an echo chamber is defined as an "environment where [people solely] encounter information or opinions [which] reflect and reinforce their own" (Hance, 2022). That environment can be seen on TikTok, which observes and analyzes every user's detail and interactions, and any of those interactions on the app result in immediate input to the algorithm. These algorithms are tailored to individuals' For You Pages and put users in a loophole of personal echo chambers. These echo chambers form algorithms that are encompassed and fueled by confirmation bias which favors information endorsing existing beliefs, making it harder to see or consider opposing perspectives and viewpoints (Hance, 2022). Engaging with user generated content (UGC), such as a video, hashtags, audio, or creators, will inform the recommendation algorithms and produce similar content related to that engagement. Despite TikTok posting public transparency reports for their users, the spread of discourse concerning digital freedom and internet freedom of speech needs to be fully transparent, especially regarding the disputable content across TikTok (Cowls, 2020). This report highlights how the subjectivity of posts seen on TikTok has shown contrary to exactly the types of posts that get moderated and removed.

Disputable Content: Disinformation and Misinformation

Listed in the community guidelines under the section 'integrity and authenticity', misinformation content on TikTok is prohibited. That section includes topics ranging from civic processes to public health and safety that could mislead the TikTok community (TikTok, n.d.). Within 'integrity and authenticity' is a subsection called 'harmful misinformation', inaccurate content includes, "misinformation inciting hate/prejudice, medical information, conspiratorial content, elections and other civic processes, and digital forgeries" (TikTok, n.d.). Despite this definition, there is no specification as to how processes such as inaccurate election data, digital forgeries, and medical misinformation are reviewed. TikTok does not list any severe consequences for users who violate this policy either. The only punitive action outlined for a misinformation violation is that TikTok will remove the content and any

associated accounts (TikTok, n.d.). To add, TikTok only addresses misinformation and does not have a section on disinformation.

The vague nature of TikTok's harmful misinformation policy has enabled this kind of content to be disseminated across the platform on many occasions, however this issue was particularly evident during the COVID-19 pandemic. During the pandemic's start, many posts related to anti-vaccine sentiment and vaccination myths circulated on the platform (Zadrozny, 2021). These posts were altered with discouraging images, sounds, and personal beliefs that generated rapid views and widely spread false information. A New York Times article addressing the spread of COVID-19 misinformation reported that TikTok claimed to have removed more than 250,000 videos relating to COVID-19 (Hsu, 2022). However, this data represents only a tiny percentage of the millions of posts that the platform encounters every day (Hsu, 2022). Hashtags were another way misinformation content gained traction and views on the platform. Some of the most notable COVID-19 posts contained hashtags that would appear when searching 'COVID vaccine' in the TikTok search bar, such as "COVID-19," "Coronavirus," "GoAwayCorona," "GoAwayCoronaChallenge," "COVID vaccine exposed," and "COVID vaccine injury" (Li, 2021) (Hsu, 2022). Often, the hashtags contained substituted words (i.e., pandemic became "panoramic" and "panorama") which could fool computer AI and prevent it from identifying correct COVID-19 posts. Substituted words allow irrelevant COVID information videos to be taken down as AI will define those words with their actual meaning.

As a very influential platform, many people rely on TikTok for information, and TikTok's insufficient response to misinformation poses a risk to its users. There is a gap in TikTok's approach to addressing misleading information, as can be seen in the proliferation of COVID-19 posts discussed above. COVID-19 misinformation posts are being shared and reacted to, drawing in the consequences of COVID-19 misinformation deriving from a feedback chain of anti-vaccine narratives (Zadrozny, 2021). Even users who sought general knowledge on things such as home remedies or how to avoid COVID-19 would eventually be shown a video that was not entirely fact-checked (Grierson, 2021).

Although TikTok does partner with fact-checking companies (ascribed through the International Fact-Checking Network Code of Principles) who can review content in over 30 languages, as well as collaborate with Content Advisory Council, researchers, civil society organizations, and media literacy experts to help combat misinformation posts the time to allocate the misinformation post, plus reviewing and fact-checking, allows those misinformation posts to continue to reach its target audience (Keenan, 2022). Even with the platform highlighting how the app cooperates with public health

experts to create an information hub that provides users with authoritative COVID-19 information, such as the World Health Organization (WHO), these videos are often overshadowed by manipulated COVID misinformation media posts. The COVID-19 example demonstrates how even with resources implemented by TikTok to counter misinformation, these resources for content moderation show that when it comes to diluting potential real-life implications arising from the massive spread of misinformation posts, the platform cannot sway away from seeing similar disputable ideas esteeming from previous misinformation videos that users see.

Disputable Content: Hate Speech

The growing debate around various filters, effects, and editing features on TikTok to produce creative content has brought up the issue of normalizing racist content. TikTok's hate speech policy falls under the 'hateful behavior' section. This section states that the platform does not permit any content that contains and pertains to hate speech or is associated with hateful behavior (TikTok, n.d.). This section has sub-categories on hateful ideology, misleading, and infringing content. TikTok defines hate speech as content that attacks or threatens topics such as race, ethnicity, sexual orientation, gender, and gender identity (TikTok, n.d.). If posts match these definitions, TikTok will ban users who engage in hateful behavior or accounts that promote hate speech. The issue is that, despite the inclusion of definitions and consequences, TikTok fails to break down or expand upon these definitions to give concrete examples of violations constituting posts that engage in or post hate speech.

The subjectivity of TikTok's guidelines allows for any posts to avoid categorization as hate speech, which has led to the issue of videos surfacing that disguise hate speech as humor. These posts are known to be 'manipulated media' content. Manipulated media content highlights one way that TikTok fails to regulate the 'TikTok language' discourse imposed by creators. Expanding on previous mentions of substituted words, this 'TikTok language' was similarly used as a substitution so that posts would not be flagged down as hateful behavior. During the pandemic, these practices were used to disguise xenophobic comments and anti-Asian sentiment in the form of humor, such as memes about "yellow peril" narratives contribute to Asians as the main cause of infection and blaming it on how society now has to live in a "panorama" (Matamoros-Fernández, 2022; Delkic, 2022). With creators using metaphors and analogies to get around community guidelines, TikTok's AI moderation has struggled to catch references in posts that portray hateful and violent content.

TikTok is aware of rising manipulative media content on the platform, such that in its policy, digital forgeries are deemed as manipulated media posts that distort truths of events to cause harm (TikTok, n.d.). Manipulative posts, including audio, pictures, and text containing metaphors, analogies, and juxtaposition, have also promoted hateful behaviors seen as transphobic and anti-LGBTQ. A study from Media Matters examined instances where users were mocking and degrading trans people while also spreading claims that there are only two genders—messages which mainly came from far-right groups, White nationalists, and misogynistic accounts (Little, 2021).

Additionally, Media Matters reported that the posts gained hundreds of thousands of views, but only a small portion of those videos were taken down. Transphobic videos on the platform violate TikTok's definition of hateful behavior, especially regarding gender and gender identity, but are slow to be removed. Consequently, TikTok has not created a safe space for all users as dangerous and hateful ideas like transphobia and anti-LGBTQ sentiments are spread and aggrandized on the platform. Overall, TikTok's efforts for addressing hate speech fall short of enforcing accountability for users who promote or partake in spreading these ideas, as well as addressing the highly nuanced and contextual issues of users reappropriating terminology to avoid violating guidelines (Han, 2021).

Disputable Content: Extremism

In addition to disseminating disinformation, misinformation, and hate speech, TikTok has also been used as a means of promoting, gaining support for, or encouraging extremism. Listed in the Community Guidelines, TikTok prohibits videos or content that "threaten or incite violence" or videos that "promote dangerous individuals or organizations" and will ban users who violate this policy (TikTok, n.d.). The unique design of TikTok (e.g., video, audio, and text) allows for content that users make and share on the app to be also reposted and shared by other users on the platform. This aspect of meme circulation on the platform has helped facilitate the sharing of far right and White supremacist memes (Richards, 2022). An instance that led to the spread of this type of extremist content on TikTok was the January 6th insurrection of the U.S. capitol. A study by the University of Illinois Chicago revealed how manipulative media content twisted the painful and traumatic realities of the capital riot into 'fun' 60-second videos (Vickery, 2021). Incorporating TikTok's fun video-editing features, users have posted clips adding popular sounds and tones to jarringly juxtaposed pictures of terrified senators hiding behind chairs (Vickery, 2021).

Users creating humorous content out of a traumatic event like the January 6th insurrection accentuates harmful real-life impacts. The humorous content can potentially target users who use TikTok as the algorithm on TikTok will repeatedly show users content related to far right and White supremacist groups, fueling users' support of these dangerous views held by these groups. With TikTok being a user generated content platform, personalized 'For You Pages' (FYP) developed by the algorithm will cater to and fit into the desires of users (Hance, 2022). Users who interact with extremist memes will continue to see similar content on their personalized FYP.

Users can report extremist videos that can surface on their FYP. For TikTok, reporting means: 1) the video will be under review to see if the post violated community guidelines, and 2) it will be reported to legal authorities if the post is found to incite any threats after review (TikTok, n.d.). TikTok has also stated that if the platform finds any extremist organizations or individuals on the app, it will ban them (TikTok, n.d.). While flagged posts will be taken down, similar extremist videos may still be on the platform. These videos can belong to private accounts and be shared outside of TikTok, allowing the post to attract people with similar extremist behavior. Research conducted by Institute for Strategic Dialogue showed videos uploaded by ISIS with a hashtag written in the Spanish language that translated to "Islamic State" became a video used to spread ISIS propaganda (O'Connor, 2021.) The ISIS propaganda videos passed the screening of AI, given how TikTok kept the video and only added a warning tag in videos promoting propaganda with, "the action in this video could result in serious injury". These warning tags applied to videos that included violence that, in theory, violated TikTok community guidelines (e.g., a video showing ISIS militants firing rockets and rifles and orchestrating suicide car bombs) (O'Connor, 2021). These videos remained on the platform for quite some time before being deleted and reported. The action to report, while helpful in detecting harmful extremist content on the platform, is still a weakness in mitigating implications that arise from using the platform to spread extremist ideas.

Twitter

Twitter debuted in mainstream use in March of 2006. In the words of its mission statement, Twitter's "purpose is to serve the public conversation" (Twitter, 2022). Today, its market capitalization, or the value of a traded company in the stock market, lands at $31.34 billion (Nasdaq, 2022) with around 368 million active monthly users, projected to drop significantly in 2023 and 2024. The drop in users may be attributed to the change in ownership to Elon Musk (Statista, 2022).

Twitter is a popular platform globally; however, countries like China, Russia, Iran, and North Korea ban their citizens from using Twitter (Time, 2022). It is a platform for free conversation, which may explain why many non-democratic nations bar its use. Although Twitter has made headlines in the past few months, in this section Elon Musk's current ownership of Twitter will only be partially discussed when necessary, concerning updated community guidelines. Twitter was a publicly traded company until it was made private during the change in ownership; however, people are still able to invest in other venture capital companies with ownership of Twitter (Munan, 2023).

The immense increase in the presence of social media in the past ten years has pushed the boundaries of freedom of expression and blurred the lines of what platforms can and should moderate. Twitter, like other major platforms, has faced significant backlash from users, investors, and government agencies for the lack of moderation of certain content, especially disputable content such as hate speech, extremism, misinformation, and disinformation. Twitter's AI moderation systems can learn biases based on user engagement which can lead to these systems' lack of detection of disputable content. The following sections will discuss the background of Twitter, its community guidelines, and the methods Twitter uses to detect and moderate disputable content on the platform.

Community Guidelines

Advertisement is the primary revenue driver for Twitter, like many other platforms. To ensure that advertisers want to continue doing business with the company, Twitter must establish community guidelines that help moderate user content (Reuters, 2023). The community guidelines outline the various categories of content that are disallowed from being shared and are subject to consequences if posted. Some examples of the categories include safety, with more specific subcategories like violence, abuse, and harassment; privacy, with sub-categories such

as non-consensual nudity; and authenticity, which includes civic integrity and synthetic and manipulated media (Twitter, n.d.). While the guidelines may be detailed and extensive, they only represent the definitive version Twitter wants to achieve for the public. Much of the content shared via the platform violates many community guidelines and is overlooked by the moderation technology. This section will discuss how Twitter succeeds and falls short in moderating disputable and highly controversial content based on its guidelines.

Considering the community guidelines of Twitter and the significant issues that challenge content moderation today, this section will discuss noteworthy examples of disputable content and how Twitter interacts with them. As established, although the three topics of focus regarding disputable content are exempt from legal regulation, Twitter and other platforms can independently moderate them by implementing rules in their community guidelines.

Disputable Content: Disinformation and Misinformation

Under the authenticity section of Twitter's community guidelines is a statement regarding synthetic and manipulated media, or any media which can be misleading, which states that users may not deceivingly share synthetic or manipulated media that can be anticipated to inflict harm (Twitter, 2022.). It also states that any media [Tweet] found to violate this guideline will be flagged, which is available for any user to see, as a measure to "help people understand their authenticity" (Twitter, 2022).

Examples of the latter statement were particularly apparent during the peak of the COVID-19 pandemic. Many people took to Twitter to share personal medical opinions on the virus's validity, the vaccination's success, or the guidelines that the general public was given to remain safe and healthy (Twitter Transparency, 2023). The domino effect of COVID-19 disinformation and misinformation in the U.S. began with a quote from President Donald Trump in April 2020, in which he stated that injecting bleach into a person's veins would have the same effect as the COVID-19 vaccine (Politico, 2021). According to Twitter guidelines, Trump's tweet sparked a barrage of tweets from those in agreement with former President Trump's misleading and deceptive media. In response, Twitter removed President Trump's tweet and other harmful posts, facilitated counter speech by labeling posts containing medical misinformation, and one step further, directed users to accurate information regarding the pandemic (Nunziato, 2022).

Twitter also updates its guidelines weekly to broaden its definition of harmful speech and what that constitutes (Twitter, n.d.). Although Twitter flagged tweets

related to the disinformation spread by Trump, studies from the Center for Disease Control showed that 4% of respondents to a survey reported gurgling diluted bleach after seeing Trump's press conference and reading several tweets with the same information (Smith-Schoenwalder, 2020). Although 4% may be a small fraction, it points to the fact that the consequences of misinformation, when not effectively moderated, are indeed real and can be dangerous.

Disputable Content: Hate Speech

Twitter's commitment to promoting public conversation allows space for all kinds of speech to persist, including dangerous speech such as hate speech. Hate speech falls under the safety tab in the community guidelines section for Twitter, which states that users may not engage in harassment or promote violence against or target an individual or group of people, which includes any race, ethnicity, religion, and sexual or gender identity (Twitter, 2022). Unlike misinformation hate speech is not allowed to remain seen "but labeled" on the platform; once detected, it is flagged and taken down. Twitter updated its software in response to a significant increase in hate speech on the platform after Elon Musk acquired the company (Brookings, 2022).

A significant instance of hate speech on Twitter was authored by Kanye West during one of his rants on the platform in which he wrote that he wished to go "death con three on Jewish people" (Twitter, 2022). This had not been the first instance of inflammatory posts by West; however, this was the first in which he directly targeted and threatened violence on an entire ethnic and religious faction of people, which violated Twitter's policy for hate speech. Historically, West spend time posting bursts of posts on Twitter and Instagram lasting about 24 hours before going radio silent again. Twitter had previously banned and allowed him back on the platform with posting privileges. Following his tweet about killing Jewish people, Twitter banned West's account after 48 hours, making him unable to publish tweets until the platform recovered his account. While this is not the first instance of hate speech dealt with by Twitter's team of moderators, it generated significant public outcry because of several factors, some of which are the following that West has on the platform and his widespread influence on society. Antisemitism has been on the rise, with 2021 being the highest year on record for reports of harassment, vandalism, and violence directed toward Jewish people based on studies by the Anti-Defamation League (Hagen, 2022). As platforms like Twitter allow space for hate speech to persist without consequence, it becomes more likely and accepted for the words to become active in mainstream society.

Disputable Content: Extremism

Social media platforms are now being utilized as spaces for extremist groups to organize, communicate with, and recruit members from anywhere in the world. In attempts to manage this rise, as of January 2023, under the Safety heading of the Twitter Rules, the platform states its zero tolerance for hateful entities like terrorist organizations, extremist groups, perpetrators of violent attacks, or individuals who may promote these acts to interact with the platform (Twitter, 2023). Social media platforms are now becoming the space for recruitment, radicalization, mobilization, and the response of violent extremist groups (DOJ, 2014). They serve as global connectors for users to gain information and communicate with people with similar urges for violent and extreme behavior. While Twitter can update its guidelines to include specific examples of this behavior with frequency, it does not put an end to occurrences of them.

In August 2015, a spike in the average following of accounts supporting ISIS occurred and went unsuspended by Twitter for 30 days (Berger, 2016). The lapse of moderation for a month is a prominent weak spot regarding Twitter's ability to remain in control of moderating disputable content on its platform. The long delay in moderating these accounts allows for followings to grow and information that can potentially cause real-world issues to spread to the masses. The micro-blogging and chat room structure of Twitter allows for cohorts of people, with both good and bad intentions, to share similar ideas and experiences that often exist under the radar of Twitter's moderation technology. These examples highlight the importance of content moderation to mitigate the real-world consequences that arise from the abuse of social media.

Facebook

Facebook is a social networking service owned by Meta Platforms where people can make online connections, post, message each other, and share content. Mark Zuckerberg, Eduardo Saverin, Dustin Moskovitz, and Chris Hughes founded Facebook in 2004. Facebook today is considered one of the most powerful social networking websites in the world, with approximately 2.9 billion active monthly users (World Population Review, n.d.). In 2021, advertisements generated over $114 billion in revenue for the company (Dixon, 2021). As of 2022, the countries to ban access to Facebook are China, Iran, North Korea, and Russia (Oremus, 2022).

In 2021, Facebook changed to Meta, two decades after Facebook was launched. Meta works to emphasize "its focus on the creation of a 'metaverse'" while disregarding its controversies regarding the spread of misinformation, hate speech, and extremism issues (Isaac, 2021). While the company has achieved becoming one of the "Big Fives," along with including Google, Apple, Amazon, and Microsoft, Meta has struggled to moderate its content and update its policies on Facebook to not only maintain its colossal user base and brand image but also evolve as moderation becomes a more prevalent issue on the internet at large.

Community Guidelines

On Meta's Transparency Center, Facebook community standards outline the content not allowed on its platform. Each section of its community standards displays specific content moderation policies with a rationale for their existence, although they are not without their controversies. To engage with this platform, users must agree with the Terms of Use, which allows Facebook to delete content, disable accounts, and take restrictive measures when needed.

Facebook's community standards outline the company's content moderation policies, but the interpretation and implementation of some have sparked controversies worldwide. There are six major sections of Facebook's community standards: violence and criminal behavior, safety, objectionable content, integrity and authenticity, respecting intellectual property, and content-related requests and decisions (Facebook, n.d.). Under each of these main sections, there are multiple subsections where Facebook further explains what and how the company is moderating content. Some sections of Facebook's community standards provide clear moderation policies with specific examples of content, while others that deal with disputable content have sparked controversies around the world.

Disputable Content: Disinformation and Misinformation

Facebook's handling of COVID-19 misinformation has been criticized, particularly regarding the spread of vaccine misinformation on the platform and its inconsistent application of fact-checking labels. This underscores the need for more effective moderation policies and collaboration with public health organizations to combat false information. Facebook has implemented policies to address the spread of misinformation on its platform. Its community guidelines state that the company recognizes that false or misleading information can cause physical harm, promote harmful health practices, interfere with elections or censuses, or be used to manipulate media (Facebook, n.d.). Facebook has partnered with third-party fact-checking companies verified by the non-partisan International Fact-Checking Network (IFCN) to review and increase the accuracy of content moderation (Meta, 2021). The company has committed to taking down COVID-19-related material that could potentially lead to harm outside of the online world, such as hate speech and false information, as they recognize that such content has the potential to cause violence or physical harm (Facebook, n.d.).

Despite Facebook's community guidelines emphasizing the company's efforts to provide public safety and authenticity, Facebook has been criticized for handling COVID-19 misinformation on the platform. For example, the World Doctors Alliance (WDA) is a controversial group of doctors and scientists who have made unsubstantiated claims about COVID-19, including misinformation about masks, social distancing, and vaccines. According to the Institute for Strategic Dialogue's (ISD) case study, the WDA has been sharing various posts, videos, and articles on Facebook to promote its views (Gallagher et al., 2021).

When a third-party fact-checker verifies information as false or misleading, Facebook applies specific labels indicating that such content has been fact-checked and found to be inaccurate (Gallagher et al., 2021). This label may include a brief description of why the content is false or misleading, as well as a link to a corresponding article or webpage with more information about the issues (Gallagher et al., 2021). The labeling means users can access accurate information with the ultimate goal of stopping misinformation from spreading. However, there was barely any application of these labels to those WDA's messages by Facebook despite their false claims (Gallagher et al., 2021). This labeling oversight demonstrates that Facebook's

dependence on third-party fact-checking organizations is not compelling enough, as they allowed false information to spread widely and potentially influence public opinion.

Moreover, it indicates that Facebook's community standards policies did not address COVID-19 misinformation. Facebook's inconsistent application of its content moderation policies was evident in Dr. Vernon Coleman's case, who uploaded over a hundred videos on the platform containing COVID-19-related conspiracy theories and anti-vaccine sentiments (Gallagher et al., 2021). Despite being the subject of 22 fact-checking articles published in eight languages, many of Coleman's videos received no fact-checking analysis by Facebook (Gallagher et al., 2021). The platform failed to uphold its community guidelines and its commitment to the public to restrict or remove content containing false information about the pandemic and COVID-19 vaccines.

Facebook is an incredibly influential platform where people often seek information about various topics, including vaccines and health information. However, the platform has been criticized for allowing the spread of vaccine misinformation, which poses a significant risk to public health. Studies have shown that people who are exposed to vaccine misinformation on Facebook are less likely to get vaccinated, which can have severe consequences for both the individual and the community (Cascini et al., 2022). Facebook's shortcomings in addressing misinformation on its platform could be improved through more effective moderation policies, better promotion of accurate information, and working with public health organizations to combat false information. Given its vast reach and influence, it becomes clear that if Facebook fails to address this issue, there can be severe consequences.

Disputable Content: Hate Speech

Facebook's prioritization of financial interests over public safety has failed to adequately address the spread of hate speech on its platform, as evidenced by internal research and the company's response to the January 6th riot. Facebook's definition of hate speech includes any speech that involves direct attacks against people using violent or dehumanizing language, harmful stereotypes, expressions of contempt, disgust, or dismissal, statements that indicate inferiority, cursing, or calls for exclusion or segregation (Facebook, n.d.). According to its community standards, Facebook uses a tiering system to categorize content and prioritize its review process (Facebook, n.d.). Under the tiering system, content considered the most severe, such as hate speech or graphic violence, is assigned to the highest priority tier and is

reviewed by a team of moderators faster than other content (Facebook, n.d.). While the company website clearly emphasizes how Facebook does not allow organizations or individuals that state a violent mission, there have been clear instances of this system failing.

When the supporters of Donald Trump stormed the U.S. Capitol on January 6th, 2021, insight from the Facebook whistleblower Frances Haugen highlighted how the company prioritizes financial growth over public safety (Suderman & Goodman, 2021). The internal documents provided by Haugen indicate that Facebook has known for years that its platform has been utilized to spread harmful content, such as hate speech, and that the company has failed to adequately address these issues (Suderman & Goodman, 2021). The internal research also explained that Facebook was aware of ways to reduce the spread of harmful content, such as political polarization, conspiracy theories, and incitements to violence, for several years. However, executives at the company have reportedly been hesitant to implement these measures out of concern for how they would be perceived by the public and investors (Timberg et al., 2021). This indicates that Facebook's priorities are more aligned with its financial interests than with its users' well-being and its consequential impacts around the world.

Facebook has responded slowly to concerns about spreading hate speech and extremism on its platform. Haugen's documents explained how Facebook allowed extremist groups to organize and recruit on the platform, contributing to real world violence and harm (Suderman & Goodman, 2021). Facebook did not effectively identify or address the communication on its platform leading up to the January 6th riot. Participants sought to stop Congress from certifying President Joe Biden's election victory. According to one of the whistleblowers, Facebook recognized the potential risks associated with the 2020 election and implemented security measures to reduce the spread of misinformation (Segal, 2021). However, many of these measures were only temporary and did not provide a long-term solution to the problem of misinformation on the platform (Segal, 2021). Facebook's methods for detecting and removing harmful content have yet to be executed competently or professionally, which has failed to take adequate measures to address the spread of harmful content. This highlights the urgent need for increased transparency and accountability on Facebook's content moderation policies and a reassessment of the balance between its corporate profits and public safety.

Disputable Content: Extremism

Despite Facebook's efforts to moderate extremist content on its platform through various measures, some criticize these actions as inadequate as having allowed extremist content and groups to persist. According to Facebook's community standards, extremist content supports or praises individuals or groups involved in violent, hateful, or harmful behavior (Facebook,n.d.). This includes any content that promotes or advocates for terrorist activity, hate groups or hate speech, mass murder, human trafficking, organized violence, or criminal activity (Facebook, n.d.). Facebook has taken several measures to moderate extremism on its platform, including removing content that violates its community standards, banning extremist groups and individuals, and partnering with third-party organizations (Facebook, n.d.). However, these efforts have been criticized for not going far enough and for allowing extremist content and groups to continue proliferating on the platform (U.S. Department of Justice, n.d.).

One notable example of this insufficiency is the case of the Wolverine Watchmen, the militia group that was responsible for planning the kidnapping of Michigan Governor Gretchen Whitmer. The Wolverine Watchmen is a right-wing, anti-government group known to promote its agenda and connect with like-minded individuals on Facebook (Witsil, 2020). Like many other extremist groups, they have utilized the platform to reach out to potential recruits through Facebook pages, groups, and ads to share news and updates about their activities and how to join the group (Beckett, 2020). These methods were similarly used to coordinate and promote the kidnapping plot, which was foiled by the FBI (Beckett, 2020). The group members shared photos of Whitmer's home and discussed plans to target law enforcement and government officials. In 2020, the group also utilized Facebook to coordinate armed protests against COVID-19 restrictions in Michigan (Beckett, 2020).

Facebook did show its effort to prevent such cases by removing several pages and groups associated with the Wolverine Watchmen for violating the company's content moderation policies on extremism and hate speech. However, it is possible that the group is still active on the platform under different names or using different tactics for its recruitment. Despite repeated warnings from advocacy groups and experts, Facebook failed to take more aggressive actions toward removing extremist content from its platform (Beckett, 2020). Ultimately, these instances reveal that the measures Facebook has taken thus far to combat extremist content online have been insufficient, and this negligence has facilitated tangible impacts.

Instagram

Kevin Systrom launched Instagram on October 6th, 2010, as a photo and video-sharing social media application. During the first active day on the market, the platform racked up over 25,000 users (Blystone, 2022). One of their primary investors, Benchmark Capital, valued the company at around $25 million only a few months after the app was made available (Blystone, 2022). By the beginning of 2012, the platform had over 27 million users and continued to grow, with new users flooding in daily as the app became more widely available (Blystone, 2022). As Instagram's popularity rose, Mark Zuckerberg offered to purchase Instagram for nearly one billion dollars in cash and stock (Blystone, 2022). Systrom accepted; however, he made a crucial provision in the sale that the platform would remain separately managed from Facebook (Blystone, 2022). By 2019, the app had accumulated 117.2 million users, with over one billion logging onto the platform monthly (Faria, 2023). Following YouTube, Instagram has become the second most downloaded app on the Apple and Android stores (Faria, 2023).

Despite its popularity, Instagram is banned in several countries: North Korea, China, Russia, Iran, Turkmenistan, and Uganda, as well as others (Faria, 2023). Countries such as Vietnam, Turkey, and Bangladesh will occasionally place temporary bans on the platform based on their current government conflicts (Faria, 2023). Over time, over 10.5 million posts per category have been removed from the platform due to violating their community guidelines (Community Standards Enforcement, n.d.).

Community Guidelines

Given the large number of users on Instagram, a high volume of content is being published on the platform, with nearly 95 million posts and videos uploaded daily (Dixon, 2022). Instagram's community guidelines play a crucial role in shaping the platform's content moderation policies, providing users with clear rules on what is and is not acceptable. Instagram's Community Guidelines are a set of rules and policies that govern the behavior of its users. The guidelines cover many topics, including safety, privacy, authenticity, and respect. Some key areas include the prohibition of hate speech, bullying, and harassment and the promotion of violence or self-harm (Community Guidelines, n.d.). Users are prohibited from sharing graphic or sexually explicit content, engaging in spam or phishing activities, selling regulated goods, and spreading misinformation (Community Guidelines n.d.). Instagram also has strict guidelines around intellectual property, requiring users to respect the copyrights and

trademarks of others (Community Guidelines n.d.). These guidelines promote a safe and positive user environment, free from harassment, hate speech, and other harmful content (Community Guidelines n.d.). When a user's content is found to violate these guidelines, Instagram may take a range of actions, including removing the content, issuing a warning, disabling the user's account, or referring the matter to law enforcement authorities (Community Standards Enforcement, n.d.).

However, these guidelines have also been criticized for causing problems regarding freedom of speech. Critics argue that the guidelines are often vague and inconsistently enforced, leading to users being censored or even banned for expressing opinions or sharing content that may not necessarily violate the guidelines but is deemed inappropriate by the platform (Jones, 2023). Some also argue that the guidelines disproportionately target specific groups or viewpoints, further restricting freedom of speech (Jones, 2023). Overall, Instagram's community guidelines were designed to foster a safe, positive, and respectful environment for its users while protecting the integrity of the platform and the rights of its users. With that being said, these guidelines serve as a polished image for their audiences but do not provide full detail of their effectiveness.

Disputable Content: Disinformation and Misinformation

Despite Instagram's efforts to combat misinformation and disinformation on its platform, the prevalence of false information continues to pose a significant challenge to its content moderation policies. While the term disinformation is not mentioned in their community guidelines, Instagram defines misinformation as any content that is false or misleading and has the potential to cause harm. Misinformation can include many topics, from public health and safety to politics and current events (“Reducing the", 2023). Amid the COVID-19 pandemic, the platform faced many issues regarding its moderation tactics and algorithmic recommendations on topics such as misinformation surrounding the vaccine, masks, and the pandemic (Hern, 2021). In August 2020, the platform's algorithmic suggestion pages, such as the "explore" page and "suggested post" section, were leading users to several different anti-vaccination and wellness influencers (Hern, 2021). Problems arose as a result of those accounts misleading people about the global pandemic, with posts such as "no pandemic," "stop getting tested," and "the vaccine is not real" (Hern, 2021). From September to November 2020, Instagram recommended 104 posts containing misinformation, or about one post a week per profile, to 15 profiles set up by the UK-based nonprofit Center for Countering Digital Hate (Bond, 2021).

Despite Instagram's efforts to combat the spread of misinformation on its platform, the sheer volume of content and the constantly evolving tactics of bad actors have made it challenging to keep pace with moderating misinformation effectively. Instagram states that it is committed to combating the spread of misinformation on its platform and has implemented various policies and strategies to address this issue ("Reducing the", 2023). These policies include fact-checking and warning labels, as well as working with independent third-party fact-checkers to verify the accuracy of the content (Instagram, n.d.). When a post is flagged as potentially false or partially false, Instagram will remove it from the "explore" and "hashtag pages" to decrease its visibility and prevent it from spreading in users' feeds and stories. At the beginning of this year, Instagram included a feedback option called "False Information," which, along with other indicators, enables them to detect and respond to potential instances of false information more effectively (Instagram n.d.). Instagram's struggle to effectively moderate and reduce the spread of misinformation on its platform highlights the need for continued improvement and innovation in content moderation policies and technologies to ensure a safer and more responsible social media environment for all.

Disputable Content: Hate Speech

The rise of hate speech on Instagram has presented a significant challenge for the platform, calling into question the effectiveness of its content moderation policies in protecting users from online abuse and discrimination. Instagram defines hate speech as any content that attacks, dehumanizes, or promotes discrimination or violence against an individual or group based on their protected characteristics (Instagram n.d.). These characteristics include race, ethnicity, national origin, religion, gender, sexual orientation, or disability. Examples of hate speech on Instagram could include derogatory language, slurs, or symbols that promote hate towards a particular group or individual (Instagram n.d.). Instagram's community guidelines prohibit hate speech, and the platform takes a strong stance against it (Community Guidelines, n.d.). The platform uses machine learning algorithms to identify and flag potentially violating content for review proactively. Then human moderators assess the content to determine if it meets the criteria for hate speech (Vincent, 2019). When content is reported for containing hate speech, Instagram may remove the content or disable the user's account, depending on the severity of the violation (Community Guidelines, n.d.). When necessary, the platform has also disclosed that they work with law enforcement to see that brutal acts of hate speech are prosecuted by the local

authority (Carmen, 2021). In addition to these measures, Instagram has implemented educational resources and tools to help users recognize and avoid hate speech. More specifically, the platform guides how to report and block content that contains hate speech, as well as information on how to use Instagram's filtering and comment moderation features to prevent it from appearing in users' feeds (Instagram n.d.).

The platform has attempted to crack down on implicit forms of hate speech, for example, blackface and antisemitic tropes (Carman, 2021). Some posts the platform looks out for are racist memes, white nationalist content, and sometimes screenshots of fake news articles (Coleman, 2019). There have been some moderation policy updates stemming from the unfortunate event in the UK where three Black soccer stars, Bukayo Saka, Marcus Rashford, and Jadon Sancho, were targeted with racist abuse on Instagram after losing a match at the Men's Euro Open in 2020 (Criddle, 2021). After missing their penalty shots which cost the team the championship, their Instagram accounts were flooded with racist and abusive comments and emojis from fans (Nowill, 2021). BBC News reported comments on Saka's profile, and in minutes they received a notification saying the platform's technology "found that this comment probably does not go against our guidelines" (Criddle, 2021). After a day had passed, Bukayo Saka responded to the abuse on behalf of all three players by posting a statement on Instagram which detailed that he wished the platform (Instagram) was taking a more serious stance on preventing the hateful comments and had more urgency in taking down the violations (Nowill, 2021). Following Saka's remarks, several users continued to report the hate speech on his profile. However, no review confirmation was received, and more hateful comments had yet to be flagged or removed (Criddle, 2021).

Meta, the owner of Instagram, stated that they do not condone racist abuse on their platform (Criddle, 2021). Within days, the platform removed comments and accounts that targeted England's footballers. The company pledged to continue enforcing its community guidelines and acting against those who violate its rules. Instagram also encouraged players to use the "Hidden Words" tool to prevent abusive content from appearing in their comments or direct messages. While acknowledging this is a challenging issue requiring ongoing efforts, Instagram remains committed to ensuring its community is safe from abuse. Of the 105 accounts that were found to have engaged in racial abuse against England footballers, 88 are still active on the platform. They described this situation as falling far behind what the social network had promised to do to address the issue (Criddle, 2021). Instagram's shortcomings in moderating and reducing the spread of hate speech on its platform have serious real-world consequences, posing risks to the mental health, well-being, and safety of

individuals who may become targets of such speech, underscoring the need for continued improvement in content moderation policies and technologies.

Disputable Content: Extremism

The rise of extremist content on Instagram has prompted concerns about the platform's ability to effectively identify and remove such content, necessitating more robust measures to counter the spread of dangerous and violent ideologies. Instagram defines extremism as any content that promotes or supports terrorism, organized hate, or violent extremist ideologies (Community Guidelines, n.d.). This can include content that advocates for or glorifies acts of violence against individuals or groups and seeks to recruit others into extremist organizations or movements (Coleman, 2019). Instagram's community guidelines prohibit the promotion of terrorism or extremist ideologies on the platform, and the platform takes a strong stance against this kind of content (Community Guidelines, n.d.). Instagram identifies, and moderates' extremism similarly to the other disputable content discussed, integrating automated technology, human review, and partnerships with outside organizations. Proactive flagging is done by AI, while users can report content that also gets reviewed. Instagram's human moderators, third-party organizations, and law enforcement agencies work to identify and remove content that is determined to meet the criteria for promoting extremism (Gorwa, 2020).

The platform also participates in several initiatives to combat extremism, including the Global Internet Forum to Counter Terrorism, which aims to coordinate efforts among tech companies to identify and remove extremist content from their platforms (GIFCT, n.d.). When content promoting extremism is identified, Instagram may remove the content, disable the user's account, or report the user to law enforcement, depending on the severity of the violation (Community Standards Enforcement, n.d.). The platform also works to prevent the spread of this type of content by limiting its visibility to appear in the algorithmic users' feeds and search results (Gorwa, 2020).

An instance that highlighted the shortcomings of Instagram's extremism moderation procedures was the platform's handling of ISIS-related content in the years leading up to 2019. Despite Instagram's efforts to remove accounts and content promoting terrorism, a study by the Counter Extremism Project (CEP) found that ISIS-related content remained prevalent on the platform ("Extremist content", 2022). The study found that even when users flagged and reported content, it often remained on the platform for hours or even days before being removed. Furthermore, the CEP

found that Instagram's automated technology could have been more effective at detecting and removing extremist content. The platform relied too heavily on users to identify and report this content ("Extremist content", 2022). The study additionally identified several instances of Instagram recommending extremist content to users through its "Explore" feature, which could potentially lead users down a path of radicalization. This example illustrates social media companies' challenges in identifying and removing extremist content and the potential risks associated with poor moderation practices (Diaz, 2021). While Instagram has since taken steps to improve its moderation policies and address the extremism on the platform, the challenges of identifying and removing extremist content continue to pose a threat to the safety and well-being of its users, highlighting the need for further action and investment in new technologies to ensure a safer and more responsible social media environment for all.

European Union

The European Union, composed of 27 countries within the continent of Europe, is a transnational political and economic union that governs its member countries as a collective. The EU establishes laws and practices in which the member countries must abide by to participate in the collective political hierarchy and economy. In the past two decades, the European Union has adopted progressive content moderation policy enhancing the responsibilities of the social media platforms.

Content moderation practices within the EU are contingent upon the EU's law of freedom of expression. The Charter of Fundamental Rights of the European Union protects the freedom of EU citizens to spread information without government interference, specifically through the use of media (European Union, 2012). Therefore, users of social media platforms within the EU are protected by EU law. However, illegal content is not protected by the charter. Therefore, content moderation practices have been adapted to prevent the spread of illegal content and preserve freedom of expression.

Content moderation law within the EU has greatly evolved from its initial implementation. The basis of content moderation law was established in 2000 by the EU Directive on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce in the Internal Market (Directive on Electronic Commerce). This law constructed the legal framework of online services that operate within the EU domain, instituting requirements of company transparency, consumer information protections, and methods of conducting communication and electronic commerce (Directive (EU) 2000/31). More specifically, the legislation targets "information society services" which constitute any electronic service that is provided to an individual upon request in exchange for payment (Directive (EU) 2015/1535). One key component of the directive is the liability exemption clause, which removes the liability from companies for content created and shared by their users unless they are aware of the illegal content and fail to remove it (Directive 2000/31/EC, Article 14).

Under the Commission Recommendation on Measures to Effectively Tackle Illegal Content Online, the EU defines illegal content as material that relates to "terrorism, child sexual abuse, hate speech, or infringements of consumer protection laws" (Commission Recommendation 2018/334). In the past decade, several laws within the EU have begun to classify social media platforms as information society services, placing these companies under the jurisdiction of Directive 2000/31. This means that they are not liable for the content posted on their platforms under the liability.

Furthermore, Directive 2000/31 does not require social media companies to actively monitor their platforms for illegal content (Directive (EU) 2000,31, Article 15). Although the companies may not be held liable for the content created, the EU has increased the responsibilities of the companies to take down illegal content, report such instances to the European Commission, and institute safeguards to protect their users through new legislation (Regulation (EU) 2022/2065).

Since 2000, social media use has increased exponentially, demanding a revaluation of Directive 2000/31 in its modern application of content moderation. In response to this growth and further development of social media, the EU has established a new piece of legislation that places harsher restrictions on social media companies to take initiative beyond the previous practices of self-regulated content moderation (Regulation (EU) 2022/2065). This new piece of legislation is the Digital Services Act (DSA), which harmonized the obligations and protections allocated between social media companies and their users across the EU member states (Regulation (EU) 2022/2065). Ratified in 2022, the DSA expanded upon the framework of Directive 2000/31 to enhance the obligations and mechanisms for social media platforms regarding the removal of illegal content as defined by Directive 2000/31, standardizing content moderation laws across all EU countries (Regulation (EU) 2022/2065). The DSA places greater emphasis on companies to moderate content within their domain, improving transparency and accountability within their content screening and removal processes.

Under the DSA, once the social media platforms are made aware of illegal content on their platform, either by order of a national authority, or independent entities with expertise in detecting and removing illegal content known as "trusted flaggers", the platforms are required by law to remove such content as to maintain their liability exemption (Regulation (EU) 2022/2065). Under the new regulation, social media companies are additionally required to report criminal offenses and or illegal content to the national government and EU Commission. The reporting authority is dependent upon the size and reach of the platform; smaller social media platforms report to the country domain in which the offense was reported whereas those that reach over 10% of the population, such as Facebook and YouTube, are designated "very large online platforms" (VLOP) and placed under the jurisdiction of the EU Commission (Regulation (EU) 2022/2065). The legislation enhances external and internal auditing practices and appoints trusted flaggers to aid the content moderation process (European Commission, 2022b). Externally, once a year, social media companies must commission an audit from an independent organization to evaluate their compliance with the DSA (Regulation (EU) 2022/2065). Internally, transparency

reports are the main mechanism in which companies audit their content screening processes. Once a year, all social media companies are required to draft a transparency report containing the number of notices received regarding illegal content and disclosure of how they use algorithms in the content moderation process (Regulation (EU) 2022/2065). VLOPs are also required once a year to account for systemic risk posed by the function and use of their service, specifically regarding the potential dissemination of illegal content and negative effects in relation to violence and mental health. This risk assessment will be used as guidelines to adapt their terms of service, algorithms, and content moderation process to mitigate the present risks (Regulation (EU) 2022/2065). Additionally, VLOPS will be required upon request by the European Commission to disclose information regarding the use and functioning of their content sharing algorithms to ensure compliance with the DSA (Regulation (EU) 2022/2065). The DSA heavily improves the accountability of VLOPs and social media platforms, placing significant emphasis on protecting the privacy and security of its citizens.

The DSA places harsher restrictions than previous legislation on social media platforms that fail to adhere to its regulations. The failure to comply with the obligations set forth in the DSA will result in a fine worth 6% of the annual turnover of the service provider (Regulation (EU) 2022/2065). The DSA went into effect on November 17, 2022, with the first report on the number of active users from social media platforms and search engine providers due February 17, 2023 (Regulation (EU) 2022/2065). Since the legislation is relatively new in its implementation, these reports will establish a baseline for the requirements of transparency reports and institution of trusted flaggers within the content moderation process.

Critics of the DSA have argued that the definition of illegal content is ambiguous, with multiple possible interpretations of what constitutes hate speech and terrorism (Turillazzi et.al, 2022, p.10). Additionally, the member states of the EU have individualized legislation regarding what constitutes illegal content, proving problematic in the universal application of content moderation; some countries may designate content as harmful but not illegal, whereas others under the EU jurisdiction may label the same content as illegal. Germany, Italy and Poland have legislation that designates defamation of religion as a criminal offense, whereas Denmark and France do not (Turillazzi et.al, 2022, p.10). The lack of clear definitions and conflicting state laws in what constitutes illegal content is problematic for companies' processes of effectively interpreting and moderating content.

Furthermore, the requirements of notice-of-action mechanisms and systemic risk assessments prioritize the removal of flagged content to stay within the legality

of the DSA, rather than effectively assessing the content for illegal material. Many companies employ artificial intelligence mechanisms to assess their algorithmic content moderation, which critics have argued is unable to effectively scan for contested content because the vague definitions cause these mechanisms to over-remove (Mchangama, 2022). These requirements, which prioritize the fast removal of content, may remove legitimate content and infringe on the users' freedoms of expression.

Disputable Content: Disinformation and Misinformation

Currently, the EU does not have a specific regulation regarding the content moderation process of misinformation. Misinformation has been definitionally grouped into disinformation by the EU under the Strengthened Code of Disinformation of 2022, which is addressed below.

In 2018, the European Union Commission implemented the Code of Practice on Disinformation, a voluntary, self-regulatory set of standards that social media companies agreed to uphold. The code defines disinformation as content that is "verifiably false or misleading information" that is "created, presented and disseminated for economic gain or to intentionally deceive the public" that has the potential to cause public harm (European Commission, 2022c). The code signifies the social media platforms agreement to combat disinformation by employing various methods of verification and fact-checking. The code outlines that companies must take appropriate measures to employ verification tools, labeling methods of political paid-advertisements, and transparency reporting in political advertising (European Commission, 2022c). Additionally, the code requires annual reports commissioned by the social media companies to demonstrate their compliance with the code and improvement in content moderation practices regarding disinformation. Facebook and Twitter signed the code in 2018, with TikTok following suit in 2020 (European Commission, 2022a).

A year following the publication of the code, the social media platforms released self-assessment reports for upholding the requirements of the code. Facebook reported within the months of March and April in 2019 that the company acted on over 60,000 ads in the EU that violated the company's adopted policies regarding disinformation (European Commission, 2019a). Similarly, Twitter established within its advertising policy the prohibition of content regarding "unacceptable business practice", including "misleading, false, or unsubstantiated claims", as well as hateful content; the company reported that it rejected over 11,307 ads for violation

of such unacceptable business practices between the months of January and August of 2019 (European Commission, 2019a). The implementation of the Code of Practice of Disinformation has been adapted across several VLOPS operating within the EU, demonstrating the movement towards progressive content moderation.

The Strengthened Code of Disinformation, published in 2022, enhanced the policy recommendations embedded within the original code. The new code expands the definition of disinformation to include "misinformation, disinformation, information influence operations, and foreign interference in the information space" (European Commission,2022c). This new definition increases the scope of potential content in which social media companies need to actively search for in their moderation processes. In addition to the original code, the social media platforms are now required to actively screen for disinformation; create eligibility requirements and content review processes for content monetization; give access of services and data to independent auditors; allow users to flag content and dispute removals; and develop tools to identify content that is harmful or containing disinformation (European Commission, 2022c). In alignment with the DSA, every six months VLOPs are additionally required to commission a report on their efforts to uphold the requirements embedded in the code, as well as commission independent audits by an outside, impartial organization (European Commission, 2022c). The Strengthened Code of Disinformation mirrors the EU's enhanced responsibility of social media platforms to engage in active content moderation. The code went into effect on June 16th, 2022, with the first self-assessment reports from companies to be commissioned in June 2023.

The Strengthened Code of Disinformation currently lacks definitive definitions of misinformation and disinformation, with significant room for interpretation regarding the "potential harm" on society from this content. The vague definitions of these terms leave social media platforms as the main interpreters for content moderation, which could potentially contravene the user's freedom of expression by over-removing legal content. As the code is entirely voluntary and relies on social media platforms to self-regulate, it could potentially create uneven levels of content moderation across various platforms.

Disputable Content: Hate Speech

The EU has produced both legally and non-legally binding policies regarding the moderation of hate speech. In the Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of

criminal law, the EU defines hate speech as such that "publicly incites violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin" (Council Framework Decision 2008/913/JHA). Any content posted on social media platforms that instigates a violent act or initiates hate against a group identified by the listed demographics is considered a criminal offense and flagged for removal. According to the decision, each member state is obliged to create their own laws in which hate speech is criminalized.

The illegality of hate speech outlined in the Council framework decision was used as a baseline for the establishment for the EU Code of Conduct on Countering Illegal Hate Speech Online in 2016 (European Commission, 2022e). The Code of Conduct on Countering Illegal Hate Speech Online creates general guidelines for social media companies, namely Facebook, YouTube, and Twitter (referred to as "IT Companies"), for moderating illegal hate speech. The code of conduct outlines the agreement of IT Companies to create mechanisms for prohibiting, flagging, and removing content containing violence or hateful material (European Commission, 2022e); once notified, the companies agree to review the majority of the flagged material within 24 hours (European Commission, 2022e). The IT Companies further agree to appoint a federal contact that provides information regarding the removal of hate speech and employ civil society organizations (CSOs) to aid in the process of flagging and reviewing illegal materials. After the publication of the code, Instagram and TikTok joined the code to voluntarily self-regulate their content for hate speech in 2018 and 2020 (European Commission, 2022e).

The IT Companies that have signed on to the code have released annual reports detailing their efforts to review and remove illegal hate speech on their platforms. From 2016 to 2019, the IT Companies reported an increase from 28% of illegal content removed to 72% in 2019 (European Commission, 2019b). Moreover, the IT Companies have improved in the time taken to review and remove flagged material: in 2016, 40% of the flagged material was removed in 24 hours, whereas in 2019 companies have removed 89% within the same time frame (European Commission, 2019b). Each of the companies have also taken individualized efforts to increase the training and employment of trusted flaggers to improve the efficiency of content screening (European Commission, 2019b). The summarized report of the commission recognizes the room for improvement in the transparency reports provided by the companies regarding the lack of reported geographical distribution of material flagged for hate speech as well as the time of the review for notices from trusted flaggers (European Commission, 2019b).

Critics of the code of conduct have argued that the code lacks the prioritization of freedom of expression and due process clauses, leaving the interpretation of illegal content to the social media platforms' discretion rather than the courts (Bukovská, 2019, p.6). Moreover, the code fails to require the intuition of mechanisms to challenge the wrongful removal of legal content (Bukovská, 2019, p.6). The lack of due process clauses and ability for users to appeal wrongful removals increases the capability of social media platforms to over-remove content, infringing on freedom of expression. Member states within the EU are given legislative freedom in instituting national laws that abide by the requirements of the code, most significantly in determining the severity of hate speech that is deemed criminal under law (Bukovská, 2019, p.6). The subsequent varying national laws are a threat to the harmonization of content moderation across the EU, leading to uneven removal levels of potential hate speech by social media platforms.

Overall, the EU has pushed significant guidelines and norms for content moderation regarding hate speech, yet the rules put in place by the code of conduct are entirely voluntary and remain practices of self-regulation within the IT Companies. The legality of hate speech is ultimately decided by the member countries to institute laws, rather than overarching legislation provided by the EU to suppress the spread of hate speech across social media platforms.

Disputable Content: Extremism

Specifically targeting illegal terrorist activity, the EU adopted the Regulation on Addressing the Dissemination of Terrorist Content Online (TERREG) in 2021 to remove the capability of terrorist organizations to communicate and coordinate their activities on social media platforms. The regulation was enacted on June 7, 2022. TERREG is an expansion of the Directive (EU) 2017/541, which set out the baseline definition of terrorist content and established that terrorist content found on social media was punishable under law (Directive (EU) 2017/541). Directive (EU) 2017/541 placed loose obligations on member states to develop their own legislation regarding the removal of terrorist content. The directive defines terrorist content as such that may cause serious damage to a country or international organization, solicits the public to commit a terrorist act, glorifies terrorist activities, or provides instructions to build weapons for the purpose of initiating a terrorist attack (Regulation (EU) 2021/784). In contrast to Directive 2017/541, TERREG establishes harsher time constraints on social media platforms to report terrorist content, and institutes authorized entities appointed by the member states to issue takedown orders for illegal terrorist content. Social media

companies have one hour to respond to a removal order created by the member state's national authority, or they risk a penalty fine of "4% of the hosting service provider's global turnover of the preceding business year" (Regulation (EU) 2021/78).

Like the DSA, TERREG requires social media companies to prepare transparency reports regarding the processes of identification and removal of terrorist content, specifically when utilizing automated content moderation tools (Regulation (EU) 2021/784). Additionally, social media platforms must develop mechanisms for users to report or flag terrorist material and means to screen their users' content. TERREG pushes social media companies to take greater diligence and initiative in their screening of terrorist activity on their platforms.

Critics of TERREG stress the limited time constraints placed on social media companies to remove terrorist material once it has been reported. The limited time could potentially hinder the capability of the companies to properly analyze the content for grounds to conduct the removal order (Rojszczak, 2022). Without the institution of a compulsory review process before the removal, there is an increased potential of content being removed that does not contain terrorist content. Content published for educational or journalistic purposes may be grouped in with the flagged material and removed. Furthermore, the vague definition of terrorist activity allows for potential state actors to issue a removal order for content moderately related to national security, posing a risk to content that is not technically containing terrorist material but interpreted as threatening to the state. The combination of loose definitions and strict time frames leave social media companies extremely vulnerable to overstepping their bounds as content moderators and removing non-terrorist content.

Strengths and Shortcomings

The European Union has significant strengths in the legislative requirements of content moderation. The EU has emphasized the importance of transparency reporting by social media companies and auditing mechanisms to mitigate the potential harmful effects of exposure to content. Additionally, transparency reports and internal and external audits stop social media companies from monopolizing their power as interpreters to remove all content, protecting user's freedom of expression. They have succeeded in many ways, including:

- The requirement of social media companies to actively monitor their platforms for illegal content improves accountability for the platforms to mitigate potential risks of harmful content for their users.

- Required annual transparency reports detailing the algorithmic processes in which social media companies remove and assess flagged content ensures universal application of content moderation and improves accountability of the platforms to institute content moderation mechanisms.
- The institution of trusted flaggers enhances the capabilities of content moderators to effectively evaluate the content and assure illegal material, protecting the users' freedom of expression.
- The capability of users to dispute content removal acts as a verification measure for social media platforms to improve their content assessment algorithms and processes.

The EU has several areas of improvement within their content moderation policies, specifically with the vagueness of definitions of illegal content, unrealistic time constraints for content removal, and lack of legally binding legislation for explicitly defined terrorism and hate speech. Listed below are those areas for improvement:

- The vague definitions of illegal content, in particular terrorism and hate speech, leave social media companies as the main interpreters of content, which could potentially lead to social media companies abusing their power and moderating content according to their own conduct policies and biases.
- EU legislation relies on social media companies to voluntarily abide by self-regulatory forms of content moderation in the realms of misinformation and disinformation. This creates unreliable and asymmetrical content moderation practices and mechanisms which may produce conflicting results of removal.
- Harsh time constraints on removal of illegal content may hinder freedom of expression because the constraint does not allow thorough process of content evaluation and auditing regarding EU law and codes of conduct. Rushing content moderation may cause social media platforms to over-remove content and overstep their bounds as moderators towards censorship.

Germany

Germany has a unique interpretation of freedom of expression in comparison to other democracies. In countries like Britain, France, and the United States freedom of expression is protected more than it is in Germany. Following World War II, Germany established strict laws on hate speech that made incitement against national, religious, ethnic, or racial groups illegal (Delcker, 2020). In this context, "incitement" means the cause, formation, or enactment of violence on an individual or group. This includes the criminalization of wearing Nazi symbols and denying that the Holocaust happened, among other actions. Germany's narrowly defined parameters have empowered the government to exert more specific and enforceable restrictions on freedom of expression. Despite these stronger regulations, the emergence of social media networks has posed increasingly difficult problems relating to disputable content moderation that pushed the German government to implement new regulations to address those issues.

In 2017, Germany passed groundbreaking legislation that set the stage for regulating online content, not only in Germany but worldwide. The Network Enforcement Act, also known as NetzDG or 'Netzwerkdurchsetzungsgesetz', was created to limit the amount of illicit content posted on social media platforms by fining companies who fail to comply with the law. Companies are required to remove illegal posts after being notified through a complaint system that is available through their platform. All complaints filed and the action taken on each complaint must be recorded in a biannual report that is published for the public to review.

Prior to pioneering NetzDG, the German government created a voluntary compliance system with social media companies in 2015 to try to limit the amount of illegal content online, although the results were not what the government aimed for (Leersen et al., 2019). As the compliance was not enforceable, companies did not follow it. Thus, there was no notable change in online hate speech. To address this shortcoming, the NetzDG was passed to create an enforceable law that incentivizes compliance by placing financial liability on telemedia service providers for removing illegal content and creating biannual complaint reports (Leersen et al., 2019). Telemedia companies are defined as:

> For profit-making purposes, operate internet platforms that are designed to enable users to share any content with other users or to make such content available to the public (social networks). Platforms offering

> journalistic or editorial content, the responsibility for which lies with the service provider itself, shall not constitute social networks within the meaning of this Act. The same shall apply to platforms which are designed to enable individual communication or the dissemination of specific content (Network Enforcement Act 2019).

NetzDG effectively transformed the previous compliance from a voluntary to a mandatory law for online platforms to follow.

It is important to distinguish that NetzDG does not establish any new criminalization, rather it only makes it so these sections of the law apply to social media. Thus, when a social media company receives a complaint from a user on a platform, it is deemed illegal if it violates any section of the German Criminal Code that is referenced. The referenced sections that are relevant to the discussion on disputable content what make NetzDG enforceable are:

> 86 dissemination of propaganda material of unconstitutional organizations, 86a using symbols of unconstitutional organizations, 89a preparation of serious violent offense endangering state, 89b establishment of relations for purpose of committing serious violent offense endangering state, 90 defamation of the President of the Federation, 91 encouraging the commission of a serious violent offense endangering the state, 100a treason forgery, 111 public incitement to commit offenses, 126 breach of the public peace by threatening to commit offenses, 129a forming terrorist organizations, 129b criminal and terrorist organizations abroad; extended confiscation and deprivation, 130 incitement to hatred, 131 depictions of violence, 140 rewarding and approving of offenses, 169 defamation of religions, religious and ideological associations, 185 libel, 186 defamation, 187 slander, 241 threatening the commission of a felony, and 269 Forgery of data intended to provide proof (German Criminal Code, 2021).

These sections help to understand how a social media platform considers a post illegal when they receive a complaint.

NetzDG states that social media platforms with over two million users registered in Germany—such as Facebook, Twitter, YouTube, and TikTok – are responsible for removing "manifestly" illegal content posts within 24 hours (Network Enforcement Act, 2019). The definition "manifestly" is not stated in the NetzDG, nor

does the law give examples of what that entails. Therefore, most content falls under the following: if a user logs a complaint but that content is not explicitly illegal at first glance, the social media company then has seven days to investigate and decide whether to remove the post (Network Enforcement Act, 2019). Moreover, companies are now required to create reports biannually if there are more than 100 complaints total within a calendar year, as well as publish procedures to create more transparency in their moderation process (Network Enforcement Act, 2019). Social media companies found to be non-compliant risk facing fines of €50 million (Network Enforcement Act, 2019).

Content is brought to a company's attention via a complaint form that can be filled out by any user on the platform. Further established in the guidelines is that such complaint forms must have clear instructions for the user to fill out; when a form is submitted, companies are notified that there is potentially illegal content on their platform. All posts that have a complaint form are flagged and put under review. If a post is deemed illegal or violates the community guidelines, it is removed from the platform. Both the flagging and removal of a post are recorded in the mandatory biannual complaint report. If the flagged post is not illegal, it is no longer flagged but must still be added to the biannual report. It is unclear if a post remains flagged after a decision that its content does not violate a community guideline despite receiving a complaint. With these newly established guidelines, the NetzDG effectively sets a precedent that companies who provide a communication space are liable for content posted on their platforms should they allow hateful or illicit posts to remain visible ("Library of", n.d.).

In June 2021, an amendment was added to the NetzDG. The Act to Amend the Network Enforcement added four clarifications to the legislation: user-friendliness of complaint procedures, appeals procedure and arbitration, transparency reports, and expansion of the Federal Office of Justice (Gesley, 2021). The first aimed to make social media platforms use more user-friendly language for submitting a complaint. Users would then know how to submit a form and be notified if the complaint was officially filed, as well as if it was under review for being removed (Gesley, 2021). Secondly, if a user's post was flagged and removed, they have the ability to ask for an appeal within two weeks from the date of the removed post (Gesley, 2021). More information must be added to the reports such as if automation was used to detect illegal content, and if so, how that automation works, what training data was used, and the procedures for quality assurance (Gesley, 2021). Furthermore, reported complaints are subdivided by the length of time taken to remove the content, as well as the information on the number of appeals filed and those which were revised

(Gesley, 2021). Finally, this amendment gave the Federal Office of Justice, the government agency responsible for various legal and administrative tasks in Germany, the power to supervise compliance of The Act to Amend the Network Enforcement (Gesley, 2021). The Federal Office is the body that determines if an infringement has occurred and requests further information on implementation measures, the number of registered users in Germany, and reports on complaints received in the previous year (Gesley, 2021).

A court case was brought against Facebook in 2021 because they deleted a user's post that was determined to violate community guidelines, but the authors of the post were not notified about the removal and blocking of the user's post and account (Goujard, 2021). The court stated that Facebook must "undertake in its terms and conditions to inform the user concerned about the removal of a contribution at least retrospectively and about a to inform him in advance of the intended blocking of his user account, to inform him of the reason for this and to give him the opportunity to make a counter-statement, which will be followed by a new decision" ("Federal Court", 2021). The court concluded that Facebook did not meet these requirements. The court further stated what forms of transparency should be in a company's terms and services for a user to have their post or account properly removed if a user violates the community guidelines. Their explanation balances an individual's right to freedom of expression and a company's right of freedom to practice a profession ("Federal Court", 2021). Facebook was then forced to reinstate the posts as the users were not informed of the decision to remove their posts. The ruling of this case came one month after The Act to Amend the Network Enforcement was instated, and this ruling upheld that social media companies have to follow the user-friendliness and transparency sections of the amendment. Furthermore, this case set a significant precedent in the European Union as it demonstrated that social media companies' content moderation complaint procedure had to be easy to access and transparent for people who want to file a complaint (Goujard, 2021).

The NetzDG has been used as a model for content moderation globally, as it is one of the first pieces of legislation to attempt to provide online protection for users. As of February 2018, countries like Russia, Singapore, and the Philippines had directly cited the law, accompanied by more in the following years ("Human Rights", 2018). In addition, since NetzDG's creation, countries such as Brazil, Poland, India, Turkey, Vietnam, and Pakistan have openly stated they took the German legislation as a model (Elliot et al., 2021).

Despite NetDG's successes, it is not without its controversy. Critics like the Human Rights Watch Director of Germany believe that the law will limit the amount of

democratic debate on online forums, thus having fewer people check on politicians' policies, actions, and comments ("Human Rights", 2018). Ultimately, the next few years will determine whether the government's legislation will accomplish its content moderation goals, or whether it is the censorship bill that many critics suggest it is (Leerssen et al., 2019).

Disputable Content: Disinformation and Misinformation

Drawing on German Criminal Code, NetzDG works against those who post disinformation and misinformation[1]. As established, if a user who is registered in Germany files a complaint on a social media platform with more than two million users, that company must take the post down within 24 hours if it is explicitly illegal, such as saying the Holocaust didn't happen. If a complaint is indeterminable at first glance whether it is illegal, then the company has seven days to decide whether to block or delete the post (Network Enforcement Act, 2019).

One of the intentions of the legislators who created NetzDG was that it would work against the promulgation of false news; however, the bill never explicitly uses the words misinformation or disinformation, so the definitions are left ambiguous. Despite the fact there are no official definitions, there are certainly ways in which misinformation and disinformation are moderated. Social media companies are required to combat misinformation and disinformation by removing flagged posts, which can later be reinstated if nothing is harmful about the post. It is the social media company's responsibility to take down posts that have a complaint form if it spreads illegal false news and include those complaints in their reports. If a company doesn't take the post down and record it, they will be fined for non-compliance.

Since the start of the COVID-19 pandemic, Germany has struggled with the spread of disinformation and misinformation, particularly from right-wing groups which facilitated a rise in the number of lies spread pertaining to the virus (Kettemann, 2022). Groups such as neo-Nazis, right wing-extremists, and anti-vaccine activists joined together to protest German mandates on the prevention of spreading COVID-19 (Schultheis 2021). However, despite an increase in anti-COVID-19 mandate sentiment, only 17% of the German population believed the government measures were going too far (Schultheis, 2021). So, although there was an increase in the number of posts

[1] This uses the German Criminal Code sections 86, 89b, 90, 91, 100a, 111, 126, 130, 131, 140, 169, 185, 186, 187, 241, and 269.

containing misinformation about COVID-19, most Germans supported preventative measures for protecting the population.

As a result of Germany having stricter limitations on freedom of speech, misinformation and disinformation are less prominent than hate speech and extremism online for German legislators. Although the German government attempted to institute measures to combat misinformation and disinformation through NetzDG's implementation, there is certainly room for improvement. Germany could work towards creating an amendment for the Network Enforcement Act that would specifically define disinformation and misinformation in order to make it clear what is legal and illegal to post. Furthermore, it is currently up to a company's discretion on the extent they will moderate fake news accounts and posts. Therefore, it could be beneficial to add an amendment that would require social media companies to prevent the dissemination of misinformation and disinformation.

Disputable Content: Hate Speech

While NetzDG is applied to other types of disputable content, the foundational premise of NetzDG is to end hate speech on social media platforms. The German government holds social media companies accountable speech that occurs on their platforms by extending established criminal laws to apply to their platforms before the implementation of the Network Enforcement Act[2]. In addition to that extension, the government also formed special online hate units in order to suppress online hate speech, which are comprised of lawyers who file complaints with the police. Users who have been affected by a piece of content can petition the online hate speech unit, or other non-profit advocacy groups, to defend a lawsuit. For an individual to pursue a lawsuit, the law requires the cooperation of the social media company, the police, and a lawyer from an institution such as the online hate speech unit, to be able to provide a specific post as a piece of evidence. One notable unit is in Gottingen, Germany, whose unit is made up of six lawyers (Satariano et al., 2023). With the help of this specialized unit providing information to police in January 2023, over one hundred accounts of police executing warrants for people accused of posting online misinformation, insults, or hateful comments, pictures, or videos. The goal of these police visits was to confiscate technology to search devices for evidence that the user had posted illegal content (Satariano et al., 2023). Although this task force unit is

[2] The following sections are used in NetzDG to make the NetzDG enforceable, and the specific sections are as follows: 86a, 90, 111, 130, 169 185, 186, 187, and 241 (German Criminal Code, 2021).

small, the intention is if they make a spectacle of individuals being fined for their illegal posts, then fewer people will post about misinformation, political extremism, and hate speech (Satariano et al., 2023).

The specialized units create an interesting new dynamic in the German legal sphere as German policymakers and law enforcement are putting increasing accountability on online perpetrators, not just the social media platforms. It is critical that telemedia companies cooperate with law enforcement when an individual files a lawsuit because the post is the primary evidence for catching the people who created that illegal post. The only way to directly fine users who post illegal content is through these individual lawsuits. However, as stated earlier, companies are not obligated to file police reports or lawsuits as Google won the lawsuit against the Act to Combat Right-Wing Extremism and Hate Crimes.

Germany made a bold move by creating specialized units for receiving complaints about illegal content. In many countries, this wouldn't work because of the sentiment to protect freedom of speech. Germany sets a unique precedent where users can no longer hide behind anonymous accounts and troll individuals' profiles without consequences. The addition of hate crime units to the execution of the NetzDG act will most likely decrease the amount of hate speech online, but an issue that the government must now be aware of is the risk of over-monitoring online platforms and over-punishing users.

Disputable Content: Extremism

Another aim of NetzDG is to prevent the dissemination and assembly of extremist ideology and of individuals[3]. NetzDG its newfound enforcement of fines is directly tied to limiting the amount of online terrorist, unconstitutional, and extremist content. NetzDG is applied similarly to misinformation, disinformation, and hate speech, with the same timeline of explicitly illegal posts being removed within 24 hours, and debatably illegal posts reviewed within seven days.

The murder of a politician, the attack in Halle an der Salle, and the attack in Hanau were the three incidents that caused the German government to add an amendment to NetzDG because the original legislation wasn't enough to prevent these attacks. The murder of German politician Walter Lubcke, the face of Germany's

[3] The bill incites the parts of the German Criminal Code that relate to terrorism, extremism, and public violence by mentioning sections 86, 86a, 89a, 89b, 91, 111, 126, 129a, 129b, 130, 131, 140, and 241 to enforce companies to remove relative illegal content.

refugee policy at the time, was one of the first incidents to highlight the consequences of online hate speech. His strong support of admittance of refugees and Chancellor Angela Merkel's immigration policies led him to being a target of right-wing extremist groups and anti-immigration activists for years (Satariano et al., 2023). In 2019, Lubcke was assassinated in his home by a neo-Nazi who disagreed with Lubcke's stringent defense of Germany's open border and refugee aid policies (Deckler, 2020). This occurrence is part of a larger, more severe problem in Germany: radical far-right activists and neo-Nazis have been convening on social media platforms (Deckler, 2020). There have been numerous racially charged attacks in Germany as these groups create echo-chambers of their dangerous beliefs and feed off one another's hatred online. The second attack was in Halle an der Saale in 2019, where a man attempted to kill 51 Jewish people in a synagogue on the sacred Jewish holiday, Yom Kippur (Escritt et al., 2019). The perpetrator tried to get into the synagogue, but the door was too thick for him to get in - he instead killed a nearby man and woman. He live-streamed the attempted attack, saying the Holocaust didn't happen and feminism is the cause of the decline of the birthrate in the West (Escritt et al., 2019). Later at trial, he maintained the same sentiments regarding antisemitism and misogyny, saying he was inspired by the Christchurch massacre in New Zealand ("BBC", 2020). The final key incident was another extremist attack in Hanau in 2020 where a man killed nine people, five of which were Turkish, making it one of the deadliest killings in post-World War II Germany (Nasr et al., 2020). Days before the attack, the gunman posted online a video of racist conspiracy theories, as well a manifesto document perpetuating these racist theories, providing evidence that this was a racially charged attack against migrant Turkish Kurds (Taylor, 2020).

Since these three attacks, Germany has increased its efforts to eliminate the spread of hatred inciting violence and extremism online. On April 3, 2021, the German government added an amendment to NetzDG called the Act to Combat Right-Wing Extremism and Hate Speech. This amendment was added following three major extremist attacks in Germany. The amendment was intended to prevent the circulation of extremist ideology online by mandating social media companies notify the German Federal Criminal Police Office of all potentially illegal content deleted and blocked from their platform (Agrawal, 2022). However, Google immediately filed a lawsuit against the new legislation as the act was not aligned with the EU Commerce Law (Agrawal, 2022). Essentially, the amendment required social media companies to send all content removed due to its illegality to the police department without notifying the user, and even if its removal was still pending (Busvine, 2021). This

meant that both innocent and guilty people's data would be given to the police department without their knowledge.

On March 1, 2022, the Cologne Administrative Court ruled that the law cannot force companies to inform the Federal Criminal Police Office of illegal content because it violates the Country of Origins Principle in the European Union's Directive on Electronic Commerce (Agrawa, 2022). Those principal states than online service provider, such as a social media company, is subject to the laws and regulations of the country in which it is established (i.e., where its headquarters are located), rather than the laws and regulations of each individual country where its services are being accessed or used (EDiMA, 2020). Google's lawsuit reversed the new amendment, but the overarching NetzDG legislation is still in effect. While companies don't have to send over individual complaints unless there's a warrant, they still must produce complaints filed in the biannual report. The urgency for the bill created choppy wording and poor practice for getting the legislation passed, so it could be worthwhile for the German government to go back and add corrections to have the amendment be properly implemented.

Strengths and Shortcomings

The Network Enforcement Act has been at the forefront of content moderation policymaking. It has set up innovative requirements to require social media platforms to take more stringent steps to address disinformation and misinformation, hate speech, and extremism online. The following are some of the most admirable and successful aspects of the law:

- The transparency requirement for social media companies to publish complaint reports as specified in section 3.1.
- Formation of the specialized units to prosecute complaints because it creates public examples that the hate speech laws are enforced online and are effective.
- Germany is pushing the boundary of restricting freedom of speech in a democracy. They are testing the balance between freedom of speech and public safety.
- Forces large social media companies to comply with the German Criminal Code by putting the responsibility on the company for the content that is posted on their websites.
- Fines individuals who post illegal content online. They can no longer hide behind a screen or be anonymous in the comments, pictures, and videos they post.

Although NetzDG is an inventive piece of legislation, it does not come without criticism. The main concern is the limitation of expression. Other concerns are that NetzDG gives too much power to social media platforms for deciding what is legal and illegal online. Other animadversions are as follows:

- Social media companies determine what is considered illegal instead of German Courts.
- Opponents of NetzDG say it is a censorship bill, giving Telemedia companies the power to decide what can and cannot be talked about online.
- There is a huge concern for over-removal which is when companies delete legal content from users (Leerssen et al., 2019). This is often done to avoid fines.
- Another major concern is that the language in the bill is too broad using terms like “hate speech”, “insult”, and “defamation”. This leads the providers to have more control over these words' meanings.
- There is not a sufficient amount of time to evaluate whether a post should be taken down, particularly in the 24-hour period (Leerssen et al., 2019).
- The specialized units of lawyers to handle complaints are too small and are being flooded with complaint forms.
- There is no standard format for the reports, making the researcher’s job of comparing data more difficult.

United States

The United States government is a "federal democratic republic" that is founded on the Constitution (Clyburn, 2011,). The Constitution is the supreme law, and no law can violate its contents (Clyburn, 2011). The U.S. has two main political parties, the left leaning Democratic party and the right leaning Republican party. Within the U.S., there has been an increase of political polarization within the last couple decades that has resulted in personal hatred for politicians, increased nationalization, and a left vs. right divide that is harming how the country is run (Lee, 2021). The conflict between parties have become personal attacks (e.g., the feud between Republic Representative Majorie Taylor Greene and Democratic Representative Alexandria Ocasio Cortez) making it more difficult for parties to work together (Goldiner, 2023). The divide between parties has led to debates over different issues, particularly social media moderation.

In the United States, freedom of expression is protected by the First Amendment (The White House, n.d.). Under the First Amendment, the government is not allowed to make laws that restrict freedom of expression. However, in certain cases, expression can be limited. For instance, if someone were to make a threat or incite violence and that action is deemed to be an "imminent threat," then that speech could be suppressed ("Free Speech", 2023). In addition to incitement, there are other limitations on free speech like defamation, fraud, child pornography (Volokh, 2017). These protections and limitations of expression also translate to the online landscape.

The U.S. has held an important role in the usage and development of the internet and is now figuring out how to moderate that landscape. Today, social media plays a prominent role in the lives of Americans - over 70% of U.S. citizens have used at least one type of social media app (Cho & Gallo, 2021). In addition to the First Amendment, another piece of legislation, Section 230, has set the foundation for moderation of content, but has undergone little change since its inception despite the vast changes that have taken place within the internet landscape (Lawson, 2021).

Section 230 was originally established under the 1996 Communications Decency Act (CDA). The CDA was initially created to combat child pornography online but was eventually ruled to be an unconstitutional restriction on free speech through *Reno v. ACLU*. Despite overturning the CDA, based on the severability clause, which states that if any part of the agreement or law is found to be invalid or unenforceable, the remaining parts will still remain valid and enforceable. Due to this clause, Section 230

was found to be constitutional, and was effectively being from the rest of the CDA and kept in place (*Reno v. ACLU*, 1997). Section 230 serves as a "safe harbor," providing protection for social media companies from liability or penalty for the content that their users post in certain circumstances, which will be discussed further later on ("What is", n.d.). In its original introduction, Section 230 encouraged free speech and allowed online content moderation without companies needing to fear liability for what was being said online (Castro & Johnson, 2021). It was also created to help new internet businesses while also encouraging the regulation of online content (Department of Justice, 2021). While Section 230 does not protect interactive computer services, like social media companies, should content on their platforms violate federal criminal law like child pornography, there is currently a large grey area within Section 230 offers protection for the publication of disputable content ("Section 230", n.d.). Section 230's provisions have resulted in numerous debates over whether Section 230 needs to be changed to address disputable content given that this content has the potential to cause societal harm.

There are two main points of Section 230 which pertain to liability when moderating content online. The first provision states that no interactive computer service or user of that service will be considered a publisher of content made by another user (a third party) ("Protection for", 1996). This means that online platforms that host content posted by third parties are not considered to be a publisher of the content on their platforms, protecting them from any responsibility for what was posted (Silver & Zarkowsky, 2022). Many call this the "safe harbor" provision. The first provision additionally implies that interactive computer services are not required to do any form of moderation. However, if a platform does decide to moderate content, the second provision protects them from liability for the takedown or restriction of certain content ("Protection for," 1996). This "good faith requirement" protects online services from mistakenly removing content or removing content that can be considered "objectifiable" (Castro & Johnson, 2021). Therefore, content moderation is left to the discretion of individual platforms as they maintain the power to identify and remove "objectifiable" content while enforcing community standards however they see fit.

However, with the development of social media and its evolving uses, Section 230 is being challenged by the government. Both the Democratic and Republican parties are applying pressure to create more explicit guidelines on how to moderate what users are posting on social media platforms in order to protect against disputable content like disinformation, misinformation and extremism. For instance, President Biden told Facebook it needed to be held more responsible for spreading

false information and that Section 230 overall needs to be revoked (Lima, 2020). Other politicians on the left side of the political spectrum such as Minnesota Senator Amy Klobuchar and Virginia Senator Mark Warner agree with this sentiment and have introduced reforms such as the 2021 SAFE TECH Act to increase liability moderation online (Kern, 2022). On the other hand, there are individuals, groups like the Internet Society, and parts of the Republican party that argue that Section 230 is vital for protecting free speech, innovation, and the U.S. economy ("Section 230", n.d.). Generally, officials on both ends of the political spectrum want to reform Section 230, but the left is on the side of 'pro-moderation', or those in favor of reforming Section 230, while the right leans towards reformation that will ensure freedom of expression or an "anti-censorship" view (Draper & Neschke, 2022).

The authors of Section 230, former Representative Chris Cox and current Senator Ron Wyden stated that their reason for creating Section 230 was to enable free speech and allow for content moderation without government intervention (Castro & Johnson, 2021). In essence, Section 230 was created to enable free speech, and those opposed to removing liability protections argue that doing so would censor users and violate their First Amendment rights. The Republican party is in favor of keeping Section 230 the same and are on the "anti-censorship" side of reforms, claiming that if online providers become liable for a wider range of content, the internet will become heavily censored ("Section 230", n.d.). The argument for "anti-censorship" is built on the importance of the First Amendment. Examples of "anti-censorship" acts include Missouri Attorney General Eric Schmitt filing a lawsuit against the Biden administration alleging that they were pressuring companies to moderate certain forms of content, resulting in the accusation that the administration was controlling political speech (Klippenstein & Fang, 2022). Many Republicans, like Texas Senator Ted Cruz, claim that platforms are censoring conservative users, citing the banning of prominent conspiracy theorist and alt-right radio show host Alex Jones from Facebook (Elliott & Cameron, 2023). Those with this view are trying to ensure that reforms to Section 230 would protect political speech, therefore protecting conservative voices (Elliott & Cameron, 2023). Overall, Section 230 is important for how content moderation occurs online, and the future of content moderation depends on the reforms, or lack of reforms that will happen in the future.

Disputable Content: Disinformation and Misinformation

Despite the rise of social media as a hub for information sharing, Section 230 does not have any provisions that protect against any type of disinformation and

misinformation. Since there is no explicit law requiring online providers to moderate misinformation, as stated previously, there has been a movement to hold platforms liable for disseminating this type of content. The COVID-19 pandemic highlighted the extent to which misinformation can easily spread throughout social media. During this time, falsehoods concerning the virus circulated rapidly, affecting COVID-19 responses. Various social media operators including Facebook and Google addressed this issue by flagging or removing content to limit the spread of misinformation (Cho & Gallo, 2021). The flagging of potential misinformation provides an example on how online providers are able to use moderation to slow types of misinformation. However, because moderation is not a requirement, many online providers remain ineffective in controlling the spread of harmful misinformation, like misconceptions about vaccines and how the virus is spread, leading to Congress working to reform Section 230 to address disinformation and misinformation.

One of the laws proposed by the Democratic Party in 2021 seeking to amend Section 230 is the Health Misinformation Act. This legislation states that social media companies would lose Section 230's liability protections if their algorithms promote health misinformation (Health Misinformation Act, 2021). The act uses the Department of Health and Human Services (HHS)'s definition of misinformation, which they define as incorrect or misleading information that differs from available evidence (Murthy, 2021). While misinformation is not illegal in the U.S., by these standards, if an interactive computer service provider promotes or does not remove health misinformation, it will now be held liable for the misinformation. The Democratic party introduced this act to slow the spread of harmful misinformation. Slowing the spread of false information during a crisis like the COVID-19 pandemic can help prevent more harm. For instance, misinformation about the spread of COVID-19 was prominent during the height of the pandemic, complicating the public health response to the pandemic (Cho & Gallo, 2021). If misinformation laws were implemented, there may be less harm spread resulting in a better public health response. However, this legislation raised a lot of backlashes from the Republican party and technology companies. Technology company executives claim that altering Section 230 can harm free expression online (Reuters Staff, 2021). Although this bill has not been implemented into law, it is an example of possible changes that can come to Section 230 and that legislators are working to change it. Overall, the question of whether the government should take action to address misinformation is one that is currently being discussed.

Outside of legislative change, there have been other efforts to help control disinformation and misinformation. These include various committees, boards, and even collaboration with social media companies. In 2019, the Department of Homeland

Security (DHS) created the Foreign Influence and Interference Board to gain intelligence about disinformation created by foreign actors, including COVID-19 and election information (Klippenstein & Fang, 2022). Moving into 2020, platforms were called up by the government to process and respond to reports of misinformation and disinformation (Klippenstein & Fang, 2022). Under President Biden, work had continued to slow the spread of false information with the Misinformation, Disinformation, and Malinformation team that now also focuses on various types of information domestically, in addition to the international intelligence collected (Klippenstein & Fang, 2022). In 2022, the Disinformation Governance Board was introduced as a means of studying and coordinating best practices in combating harmful effects of disinformation. However, following intense backlash from Republicans, it was subsequently disbanded. The board faced criticism across the political spectrum as many questioned why the government should oversee determining what is correct and incorrect, even questioning if one could trust the government to make that determination (Klippenstein & Fang, 2022). Although these teams made to combat disinformation and misinformation have been put to an end, it shows the potential for greater commitment to control the spread of this content, and that there may be alternative ways to monitor disputable content other than the creation of new laws.

Disputable Content: Hate Speech

In 2021, 41% of Americans experienced online harassment ("Online Hate", 2021). Despite this, The United States has no agreed upon legal definition of what constitutes hate speech. However, as established earlier, it is generally seen as a form of expression that humiliates or incites harm/hatred against people on the basis of race, religion, and other forms of identity ("Hate Speech", 2023). The First Amendment protects hate speech, even if it becomes offensive or hurtful, and especially if it revolves around public matters ("Hate Speech", 2023). The protection of this type of speech is exemplified by the case of *Snyder v. Phelps*. This case concerned the picketing of a military funeral with signs criticizing the U.S. and U.S. soldiers, resulting in emotional distress for the father of the soldier whom the funeral was for ("Facts and", n.d.). The Supreme Court ruled that speech like this was protected since it spoke on public concerns, ultimately setting a precedent of hate speech being protected under the First Amendment ("Facts and", n.d.). However, as mentioned above, there is some speech that is not protected under the First Amendment. This speech includes imminent criminal activity that targets specific people or groups, the distribution of child pornography, and expression related to

obscene speech ("What Does", n.d.). Since parts of hate speech are protected by the First Amendment, there is little legislation regarding its moderation, resulting in the moderation of hate speech online to be minimal as well. (Hate Speech, 2023).

Section 230's provisions allow social media platforms to regulate however they see fit, since it does not require platforms to moderate anything other than illegal acts. While the government may not be able to create legislation, they can, however, pressure platforms to regulate their own content. At the threat of legislative action and public pressure, social media companies like Twitter and Facebook have started to make changes themselves. Some examples of this are seen in the form of Facebook's Oversight Board, which acts similarly to a judicial body, making appeals to the platform's moderation rules ("Online Hate", 2021). Twitter Has also started banning political ads in response to pressure from the government ("Online Hate", 2021). Although platforms are trying to police themselves, Americans are still being harassed online, resulting in 81% wanting platforms to moderate hate speech ("Online Hate", 2021). While companies can try and create change themselves, their methods are not always effective in controlling hate speech due to bias, lack of training, and even inadequate tools that result in the public wanting the government to step in.

The moderation of hate speech needs to be implemented whether that be through legislation or self-regulation by companies. What is said online can spill into the real world and cause harm. For example, Robert Bowers regularly posted on Gab, a platform popular amongst right-wing extremists, anti-Semantic and neo-Nazi comments. He then killed 11 people at a synagogue (Guynn, 2019). It is important to try and control harmful speech online so that users are protected both online and in person.

Disputable Content: Extremism

In the U.S., there is also limited legislation addressing online extremism. However, there have been many calls from both political parties to alter Section 230 regarding violence and terrorism. Legislation, Supreme Court cases, and even pressure from the government are all methods that the government has employed to try and regulate extremism online.

One proposed piece of bipartisan legislation that was introduced is the 2021 "See Something, Say Something" Online Act. The See Something, Say Something Online Act would require interactive computer services to submit a Suspicious Transmission Activity Report (STAR) to the Department of Justice (DOJ). If a user finds a post, message, comment, etc. that assists a major crime, then the platform

must report that finding to the DOJ. ("Something, Say", 2021). A "major crime" includes any federal criminal offense including extremism. This proposed bill would make companies liable for what is posted on their platform if it assists a major crime, and that post is not reported to the DOJ. While this isn't an end all be all solution, the bill does put more power into the hands of users on the platform to be able to report content and forces platforms to moderate the flagged content or else they would be held liable, and thus vulnerable to punishment.

Beyond legislation, there are two trials currently underway in the Supreme Court that will be pivotal for how terrorism is addressed on social media in the future. The proceedings for *Twitter v. Taamneh* and *Gonzalez v. Google LLC* started in February 2023. Both cases stem from similar instances where families of a terrorism victim have filed action against social media platforms claiming that it should be liable for aiding and abetting terrorism since they did not take action to remove or prevent the spread of terrorist content on their platforms ("Twitter v. Taamneh", 2022). These two cases will help determine if companies are liable for terrorist content that is posted on their platforms, which has the potential to lead to the determination that social media platforms are aiding and abetting terrorists should they not take adequate steps to remove that content off their platforms.

In the case of *Gonzalez v. Google LLC*, Reynaldo Gonzalez sued Google through the Anti-Terrorism Act. He claimed that YouTube's algorithm aided in the 2015 Paris terrorist attacks that took the life of his daughter by recommending ISIS recruitment videos to susceptible viewers (Lima, 2023). The lower courts ruled against Gonzalez, determining that Section 230 protected online providers since ISIS were the ones producing the video, not Google. However, Gonzalez is alleging that through their algorithmic content sharing, these platforms are recommending ISIS content and videos. While this is not necessarily production of the video, Gonzalez is claiming that by pushing certain videos to users, the platform does have some hand in terrorist recruitment, thus aiding in terrorism (Robertson, 2022). The Supreme Court will have to determine if the algorithms that determine video or content recommendations to users are protected under Section 230 or not. The implications of this case can, according to the policy director Prateek Waghr at the Internet Freedom Foundation in India, set precedent for the entire world (Elliott & Cameron, 2023). Waghr argues that if a change occurs in one country, it is used as reasoning to implement that same change in other countries (Elliott & Cameron, 2023). Following his logic, if Section 230 is changed drastically or even abolished, then companies will apply the same changes, not just to the U.S. market, but to markets all around the world (Elliott & Cameron, 2023). In addition, the Internet Society's CEO and president Andrew Sullivan argues

that if the court rules in a way that will weaken Section 230's liability protections, companies will be more vulnerable to lawsuits, leaving only large companies like Google and Facebook surviving (Elliott & Cameron, 2023). The outcomes of this case can affect the social media companies and users both domestically and internationally.

Unlike *Gonzalez v. Google LLCI,* which questions Section 230, *Twitter v. Taamneh* concerns the 2016 Justice Against Sponsors of Terrorism Act (JASTA). To be successfully sued under JASTA, the platform had to have knowingly aided and abetted an act of terrorism (Millhiser, 2023). The plaintiff, Mehier Taamneh, is arguing that ISIS used Twitter to promote their views and recruit members (Millhiser, 2023). The challenge with this case is the issue of how to read JASTA - narrowly or broadly. If the justices were to read JASTA narrowly, plaintiffs would have to prove the platform aided and abetted a specific attack (Millhiser, 2023). This narrow reading requires proof that a large company knew about a very specific attack, which may be harder to prove. Conversely, if the court were to read the law broadly, then the issue can result in businesses being conscious about aiding and abetting any crime and can prevent them from carrying out normal business practices (Millhiser, 2023). For instance, a business may have sold a product like a weapon to a person and then that person commits a crime with the product. The fear is that if JASTA is read too broadly, businesses can be held liable for aiding in a crime since it was committed with their product. The same logic can be applied to social media platforms as well, effectively impacting Section 230's liability protections. Thus, a broad reading can lead to increased surveillance of customers, in addition to making business harder to operate (Millhiser, 2023). The outcome of this both cases will ultimately determine platforms are held liable for terrorist content online, putting Section 230's liability into question.

In addition to laws and court rulings, the government can pressure social media companies to regulate disputable content. For example, at the pressure of Congress and the UN, Facebook's Dangerous Individuals and Organizations (DIO) list was created. The DIO is a list of individuals and organizations that have a history of extremist activities on Facebook, and are subsequently banned from the platform (Biddle, 2021). While this addition to Facebook's Community Standards saw restrictions on certain groups, it can harm marginalized communities, as it targets aspects like religion (Biddle, 2021). In fact, much of online counterterrorism is left to the platforms and other outside experts, and in 2016 the "Madison Valley Wood" project was established (Clifford, 2021). Here, social media companies like Twitter, Google, and Facebook met with government officials who encouraged platforms to help with counterterrorism (Clifford, 2021). Due to the companies' morals and incentives from the government

(private and public partnerships) the 'Big Four' – Google, Microsoft, Facebook, and Twitter – started to remove accounts from their platforms that promoted terrorist activity. This collaboration resulted in the 2017 creation of the Global Internet Forum to Counter terrorism (GIFCT) (Clifford, 2021). This group of companies was now able to work together and manage their platforms regarding extremist content. Since legislation has been hard to pass, the government has started to outsource ways to regulate disputable content, which has seen some success.

Although there have been multiple proposed reforms regarding hate crimes, terrorism, and other harmful acts online, Section 230 has not been changed in any meaningful capacity. However, with Supreme Court cases and more legislation being proposed, change needs to happen. However, how that change will occur is still contested.

Strengths and Shortcomings

The U.S. has tried to control disputable content in various ways and has seen some successes:

- Section 230 protects interactive computer services and other users from being held liable for what is posted on their platforms and thus preventing large lawsuits against these companies.
- Section 230 protects freedom of expression and speech online.
- Collaboration with social media companies has resulted in some successes.

However, Section 230 does not address any specific issues regarding content moderation except for federal crimes like sex trafficking and copyright. And even when reform does occur, figuring out how to best balance regulation and free speech is a topic of debate because violating the First Amendment renders any reform non-applicable. For example, the 2018 Flight Online Sex Trafficking Act and the Stop Enabling Sex Traffickers Act (FOSTA-SESTA), was signed into law by President Trump. FOSTA-SESTA was introduced so that platforms can be held liable for unlawfully promoting trafficking and prostitution ("Allow States", 2017). This is one of the only amendments that has removed Section 230's immunity protection for interactive computer services. However, this change has come with many complaints from citizens. They argue that this change is an infringement on the First Amendment since it takes away a space from sex workers to promote and discuss their work, showing how reforms toe a fine line between content moderation and censorship (Juecic, 2022). Although FOSTA-SESTA has not been ruled unconstitutional, the uproar caused by the

amendment does not bode well for future reforms, as it only pertains to explicitly illegal content. The likelihood of disputable content is therefore more likely to be difficult when there is such controversy over FOSTA-SESTA.

This leaves many gaps in legislation and fails to attribute any responsibility to online providers in terms of the moderation of their content. As discussed, there have been many reforms introduced and multiple challenges in courts, especially regarding content such as misinformation and extremism. Despite efforts, there have been shortcomings of how the U.S. is running content moderation:

- Section 230 does not address extremism, dis/misinformation, and other forms of disputable content.
- Little to no reform to change Section 230 even when politicians on both sides of the aisle want change.

Taiwan

Taiwan, officially known as the Republic of China (ROC), is an island situated southeast of the Asian continent. Before its recognition as a democracy in 1987, Taiwan was controlled under martial law during a period of authoritarianism where all public communications were strictly controlled by the government (Southerly, 1987). Following the end of authoritarianism was a transition into democracy, where the people valued freedom of speech highly. Freedom of speech is the key to Taiwan's democratic transformation and is an important core value in the further development of democracy in the country. The Taiwanese government sought to protect this right, as observed in its strong protection of civil liberties ("Taiwan: Freedom", n.d.). In order to highlight the meaning and value of freedom of speech, in 2016 the Executive Yuan also established April 7 as 'Freedom of Speech Day' ("Important Policies", 2017).

As speech is protected in public the internet is similarly free. That has not come without issues, leading to Taiwan seeking to moderate online content ("Taiwan: Freedom", n.d.). Content moderation is an increasingly important sector as more and more people rely on social media for information. In fact, when it comes to using social media as a primary news source, 45% of people in Taiwan use Facebook, and another 40% use YouTube for news (Lin, 2022). However, as censorship is rarely practiced and the country lacks specific laws pertaining to content moderation, transparency, and reliability, there are concerns over the lack of oversight for government and law enforcement removal requests ("Taiwan: Freedom", n.d.).

Without a specific law requiring social media platforms to regulate content, the National Communication Commission (NCC), the highest authority of information transmission in Taiwan for making the communications and media industry administratively neutral, announced the draft of the Digital Intermediary Services Act (DISA) in 2022 (Nagra, 2022). While it is still in the works, the DISA would govern and outline the responsibilities of online intermediaries such as social media platforms, internet service providers, and e-commerce websites ("General Description", 2022). The act offers a framework for safeguarding the rights of both intermediaries and users by establishing a safe harbor mechanism, as well as requiring that platforms produce regular transparency reports about content moderation ("General Description", 2022). The DISA would further require intermediaries to delete any content that is illegal, or activities such as services related to information that violates the law when it is reported to them or brought to their knowledge. However,

it does state that intermediaries are occasionally exempt from responsibility for user-generated content ("General Description", 2022).

The draft of the DISA was sent back to the NCC and has yet to be passed by the Legislative Yuan, the unicameral legislature of Taiwan in charge of passing laws and supervising the executive branch. There are several reasons why the DISA has not been able to be established successfully thus far. First, Article 18 in the draft states that if a digital intermediary service provider believes that the information being transmitted or stored at the user's request is untrue, illegal, or violates the law, the provider may ask the court for an information restraining order and/or issue a temporary warning to the user about the information. ("General Description", 2022). The information restraining order is a court order compelling a provider of digital intermediary services to delete, limit access to, or take other steps to strike a balance between the right to free speech and public interest ("General Description", 2022). The Legislative Yuan is concerned that the definition of illegal content is unclear, and the role of each agency has not been investigated and discussed with the Judicial Yuan, Taiwan's highest court that interprets the country's laws and the Constitution, since both must collaborate to institute the criteria and definition. Second, the court has not yet ruled against misinformation because it may be too difficult if the authorities first request the addition of a warning label since there is no legal definition and criteria for misinformation. Lastly, because social media platforms are aware of the provisions related to illegal content, it may result in a tendency to delete large amounts of content to avoid being penalized; although the draft does not currently include an analysis of the impact on implications pertaining to freedom of speech (Zhou, 2022).

After holding various public draft briefings with relevant industries, civic groups, experts, and academics, the NCC said it will keep in touch with all sectors to foster agreement as any internet-related legislation must fully take into account the viewpoints of numerous stakeholders, otherwise, the rights of a party will be negatively affected or the rights of another party could be infringed upon ("NCC collects", 2022). However, the DISA is not the only legislation to deal with content moderation specifically, there are still other related methods that the government is taking to work on these issues while the DISA is being drafted and worked through.

Disputable Content: Disinformation and Misinformation

Currently, Taiwan does not have specific regulations for misinformation on social media platforms, although they have instituted policies to regulate

disinformation. As Taiwan has enjoyed the freedom of speech following its prolonged period of authoritarianism, they have also come to discover the various problems that arise with "the right to the freedom of public information" (Rickards et al., 2019). With fake news posing a serious issue, and creating problems regarding national security, the Taiwanese government has been rapidly implementing legislation in order to combat the rise of disinformation on social media. Three pieces of legislation have been amended or created and passed by the Legislative Yuan to address these concerns.

In 1995, the government promulgated the Presidential and Vice-Presidential Election and Recall Act, which states that anyone spreading rumors by text, picture, videotape, or other methods shall be condemned to fixed-term imprisonment of no more than five years if the election is affected and there are damages to the public (Blanchette et al., 2021). The law is a crucial part of Taiwan's democratic system, aiding in ensuring the fairness and integrity of presidential elections while allowing for accountability in the event that the President or Vice President fails to perform their duties (“Information about”, 2020). In this act, the term "rumor" or "false" refers to a "fabricated word" or "fiction" whose content is intentionally fabricated; the term "damage" refers to harm brought on by another person's influence over the outcome of a presidential election (“Publications”, 2012).

In response to COVID-19, Taiwan has been actively combining and proposing various measures, particularly in the enactment of the Special Act for Prevention, Relief and Revitalization Measures for Severe Pneumonia with Novel Pathogens, to give the government access to pertinent measures for epidemic prevention, treatment, and planning (“Does spreading”, 2020). Under Article 14 of this act, anyone who spreads false information or rumors about infectious disease epidemics that harm the general public or others will be penalized (Blanchette et al., 2021). The penalties specified by the regulation include fixed-term jail, detention, and other punishments, in addition to fines of up to NT$ 3,000,000 (Blanchette et al., 2021). Although these regulations are able to punish people who spread disinformation on the internet, they are not capable of prohibiting people from disseminating disinformation on social media platforms, nor do they require platforms to moderate and delete disputable content because of the lack of content moderation law (“General Description”, 2022).

These amendments aren’t the only protections in place that combat disinformation, as the government also works with the Taiwan FactCheck Center (TFC). The TFC is an NGO that performs fact-checking of information related to public affairs. The center uses representative fact-checking systems from overseas to remain neutral and avoid being affected by other factors during the process

(Verification Criteria, 2022). The organization collectively decides with the internal members on the 'to-be-checked' items after a meeting, and the fact-checking report will be verified by at least three checkers who work in TFC before publication. Furthermore, the information is made public so that people are able to self-check or provide new information.

Disinformation is taken seriously by the government, demonstrated by their efforts to create methods for stopping the spread of disinformation by instituting DISA and cooperating with the TFC to filter the disinformation. However, the existing legal system does not clearly distinguish between malicious disinformation and accidental or unintentional misinformation. For example, Article 14 of the Special Act for Prevention, Relief and Revitalization Measures for Severe Pneumonia with Novel Pathogens and Article 90 of the Presidential and Vice Presidential Election and Recall Act both established fines for disseminating rumors or false information that will harm the public. An important distinction made by both articles explains that if the content is intentionally fabricated then there is reasonable suspicion, whereas if it is disseminated or distributed under the mistaken belief that it is true then there is no intent to violate the law (Guo, 2012). However, it is difficult to prove that a person subjectively had the intention of knowingly and maliciously disseminating false facts to the public in a court of law, which is a key criterion to judge the crime. Without uniform criteria, the judges' definitions of intentionally spreading disinformation or standard of judgment are different, it is extremely difficult for judges to determine whether the defendant violated the law. Additionally, in order to meet Taiwan's strict legal definition of disinformation, it must meet three conditions: being fake, maliciously motivated, and creating harm to individuals, organizations, or social order, which might be difficult to prove and further hold people accountable. Although the law is not targeting social media platforms, it can still help moderate content by regulating users' inappropriate comments on the platforms and indirectly achieve the purpose of decreasing the dissemination of disinformation on social media.

Disputable Content: Hate Speech

Taiwan's legal system still notably lacks any regulations specifically governing hate speech because people remain committed to ensuring freedom of speech, making it very difficult for the government to institute policies to fight hate speech. Furthermore, Taiwan is a very small country with a smaller and more homogenous demographic compared to other countries, so although the law does not specifically pertain to hate speech, there are still some parts of the law related to it.

General references to hate speech can be found within some Taiwanese legislation. In order to maintain public order and ensure social peace, Taiwan implemented the Social Order Maintenance Act in 1991. Article 63.5 of the act stipulates that "anyone who spreads rumors sufficient to affect the public peace" shall be punished (Social Order Maintenance Act, 1991). The term "rumor" refers to fabrication without factual basis and without a reason, and whoever disseminates rumors will be fined a maximum of NT$30,000 or detention of up to three days (Social Order Maintenance Act, 1991). Another policy related to hate speech is the Civil Servants Election and Recall Act, which was amended in 2020 to include punishment of imprisonment for five years for those who "spread false information through text, images, videos, audios such that it causes individuals or the public to suffer a loss" (Blanchette et al., 2021). However, the main goal of this amendment was to prevent obstructing both election and recall, rather than hate speech specifically. Because these two amendments pertain more to disinformation, there are gaps in laws that moderate hate speech at large. Future laws may address this gap, or laws being drafted could extend to regulate hate speech such as the DISA bill.

Like the disinformation section, it is extremely difficult for the court to speculate whether a person had the intention of spreading rumors or not, especially considering that no law directly regulates hate speech. Moreover, many have brought forward the criticism that because Taiwan protects freedom of speech, hate speech in most situations has no legal liability and is therefore not able to be regulated. Civil society has been arguing that the government should not only protect freedom of speech, but not allow hate speech to violate human rights. Currently, the better way to reduce hate speech disseminated on social media, in most cases for Taiwan, is to rely on self-restraint, and be aware of what is being commented on or spread across platforms (Yang, 2020). However, if the government were to establish stronger legislation defining and/or regulating hate speech in the future, this could then serve to strike a balance between freedom of speech and hate speech.

Disputable Content: Extremism

In Taiwan, like hate speech, there is no law specifically targeting extremist content, but terrorist activities have been monitored under the Communication Security and Surveillance Law amended in 2018. Article 7 of this legislation specifies that in order to prevent national security from being jeopardized and to maintain social order, law enforcement agencies may issue an interception warrant to monitor communications in order to collect intelligence from foreign powers or overseas

hostile forces ("The Communication", n.d.). The government defines 'foreign powers' or 'overseas hostile forces' as an organization whose purpose is to engage in international or cross-border terrorist activities, and 'communication' as sending, saving, transferring, or receiving symbols, words, photos, audio, or other sorts of information ("The Communication", n.d.). If a person who has a Taiwanese address has communication that may endanger national security, Taiwanese authorities can request an interception warrant from the specialized judge of the High Court at the location of the national intelligence agency to surveil communication activity. The information obtained must only be utilized for intelligence gathering for national security, but if any illegal contents are discovered then the information must be turned over to the Judicial Police, Judicial Yuan, or Military Judiciary for processing in compliance with the law.

The government also promulgated the Anti-Infiltration Law in 2020 to protect Taiwan from infiltration and interference from foreign hostile forces, to maintain social stability, and to safeguard national security (Formulating Anti-Infiltration Law, 2013). The government defines 'foreign hostile forces' as nations, political entities, or organizations that conflict with Taiwan or that support violent tactics to jeopardize Taiwan's sovereignty (Formulating Anti-Infiltration Law, 2013). Article 3 of the law stipulates that no person may receive instructions from foreign hostile forces or be given authority to participate in referendum-related activities (Blanchette et al., 2021). Further, those who violate the provisions risk being sentenced to a fixed-term imprisonment of no more than five years and fined no more than NT$10 million (Blanchette et al., 2021). In this way, the government can prevent foreign hostile forces from entering government organizations and affecting Taiwan's national safety. Historically, the Taiwanese government has not yet successfully legislated laws related specifically to terrorism. With the lack of specific laws, it is almost impossible for the government to require social media platforms to remove content related to terrorism. While Taiwan's current legislation pertains more specifically to the surveillance of terrorist communication rather than regulating terrorism content, there may be room for regulation to expand and moderate extremist content in the future. However, for now, the only moderation method related to terrorism the government is able to conduct is surveillance.

A major criticism of the Communication Security and Surveillance Law is that the information that can be obtained by the police through a warrant includes both 'communication recorder' and 'communication user information', but 'communication content; is not included. Here, a 'communication recorder' refers to telecommunication numbers, communication time, usage IP, mailbox, or location information generated by

the telecommunications system, the communication user information means the user's name, ID number, and address, and the communication content means any content that has been communicated (Yang, 2022). However, in the past, police agencies have often requested platforms to provide 'communication content' by means of a warrant to detect crime and monitor terrorist activities in which privacy information may be compromised (Yang, 2022). The Communications Security and Surveillance Act prohibits the retrieval of material from communications that have "ended in the past," regardless of whether that content has been transformed into a file such as a conversation record or fax picture, or kept on a computer or server (Yang, 2022). This makes it very difficult for authorities to monitor and moderate extremist content without the permission of the law. Taiwan has devoted itself to preventing terrorist activities from invading the country, but Taiwan also has the opportunity to institute better laws for transparency on content and communications and decrease the spread of terrorism.

Strengths and Shortcomings

Although Taiwan has not yet passed a law specifically targeting content moderation, they still have some significant successes in terms of managing the internet world. The government has been working on strengthening other related laws and cooperating with third parties to prevent disputable content from spreading on the internet without infringing on people's freedom of speech. They have seen varying successes:

- The institution of law regulating rumors and false sayings aims to prohibit people from spreading illegal content on the internet by fining and imprisoning them.
- By collaborating with Taiwan FactCheck Center, government and social media platforms are able to provide more reliable information to people and decrease the chances of being misleading.
- The mechanism for supervising the communication content of terrorist groups helps reduce the chances of spreading terrorism information and potential terrorist activities.

However, there is still room for improvement in content moderation for Taiwan. It has been extremely hard for the government to enforce the moderation of disputable content on social media platforms without a specific law regulating it. Under these circumstances, the government is not able to

protect public safety and prevent users from disseminating disputable content over the internet. Despite its successes, it falls short in a few areas:

- The lack of a content moderation law, beyond drafted legislation, leads to governments not being able to require social media platforms to moderate disputable content or delete illegal content.
- Some existing definitions and criteria are ambiguous, particularly the word intention, making it hard for judges to determine whether a person had the intention to violate a law. This may serve as an obstacle to effectively prosecuting people.
- Taiwan is a country which considers freedom of speech a significant and valuable right, which is why the government has a more difficult time regulating disputable content since they must consider whether a law will infringe on people's rights.

Brazil

The 1988 Brazilian Federal Constitution protects freedom of speech and press, as well as guarantees broad access to information from many sources (UNESCO Brasilia, n.d.). Free speech is not absolute as those who act abusively and cause damages to others may be held criminally responsible for their actions (Galperin, 2014). Laws such as Brazil's Racism Law and Anti-Terrorism Laws create boundaries to Brazilians' freedom of expression online. Brazil follows a civil code system—court decisions do not establish precedent and are primarily dependent on existing legislation. Over the past few decades, the nation has taken an active position in democratically molding internet content moderation policy, passing landmark legislation, and challenging large corporations in court. Brazil, especially with its contested political landscape, is an important blueprint for democracies to study as it addresses online misinformation, extremism, and hate speech.

Online intermediaries are serviced platforms that companies provide to host user-generated content online. Brazil's landmark online intermediary policy Marco Civil da Internet (Law 12.965/2014) was drafted from 2003 – 2010. Initially conceived by academic Ronaldo Lemos, various members of academia and civil society contributed to its creation (Souza et al., 2017). Passed on April 23rd, 2014, Marco Civil da Internet protects intermediaries from liability for content posted by users, except in cases where a court order is issued. Marco Civil da Internet is similar in nature to that of the United States' Communications Decency Act Section 230, which removes the liability of online intermediaries for user generated content. The removal of this liability highlights Brazil's original preference of leaving content-policing to corporations, while deferring to civil litigation for user content disputes under the Marco Civil. By affording limited liability to online platforms to moderate content and protecting conditional free speech for users, the Marco Civil serves as a broad legislative foundation for digital rights, but also depends on other statutes to punish potential criminal online activity.

Article IV of Title II Chapter I also broadly prohibits anonymity, which has been used to remove the social app Secret from application stores for allowing anonymous messaging (Galperin, 2014). While Secret seems to be an isolated incident in Brazil's online space, freedom of expression and prohibition of anonymity seem to contend with each other as users are guaranteed data privacy through the Marco Civil yet are expected to tie their identity to their online activity (Rodriguez, 2015). Anonymity can be used to shield harmful actors from criminal punishments, but also may be

necessary to shield those of marginalized identities in online spaces where they fear harassment due to their identity (Galperin, 2014). The question of anonymity illustrates the challenges that Brazil currently faces in balancing free expression of political ideas while trying to combat hate speech tied to identity politics that heightens around election cycles.

Disputable Content: Disinformation and Misinformation

Disinformation and misinformation are not legally defined in Brazilian law. The lack of legal definition creates a grey area around whether what users post constitutes false speech, and on what grounds platforms should remove misleading content. Brazil has recently contended with a large influx of contentious political and health related online speech following the past three presidential elections and the COVID-19 pandemic. Instead of facing scrutiny under specific misinformation-oriented laws, users' potential infractions are subject to criminal codes outlawing libel, defamation, and disrupting elections (ITS Rio, 2021). Brazil's Marco Civil is not meant to detail what users are not allowed to post online, but rather that they are broadly protected in what they can say. The Marco Civil leaves the takedowns of misleading content up to the discretion of platforms, who also do not face liability for the decisions they make. The COVID-19 pandemic and previous election cycles have tested Brazil's internet liability framework and shown it is not fully capable of preventing health-related misinformation from spreading online.

Harmful misleading narratives about the COVID-19 pandemic have been spread on social media platforms in Brazil due to a lack of accountability mechanisms for users who post them, and for platforms who fail to remove them. Highlighting its struggle to appropriately address the pandemic, Brazil has recorded nearly 37 million cases over the three years since the pandemic began (Worldometer, 2023). Its leader at the time, President Bolsonaro, was well known for spreading misinformation about isolation procedures, cures, and the potency of the virus on social media sites like Twitter and Facebook (Ricard, 2021). In turn, his supporters shared said messaging, leading to a narrative that downplayed the severity of the virus's effects. Despite many platforms taking action to remove some of Bolsonaro's claims, mitigating the impacts proved more challenging due to the extensive size of his audience (Wagner, 2020). Bolsanaro's messaging around COVID-19 highlights that social media sites in Brazil are vulnerable to false, populist messages aimed at garnering a sense of collective identity that are difficult to counteract under slow moving judicial processes necessitated by the Marco Civil.

Political misinformation has also tested the limits of the Marco Civil. In 2018, pro-Bolsonaro actors spent almost $12 million Brazilian reais to send politically charged messages to illegally obtained lists of users. The messages targeted the credibility of political opposition (Mello, 2018). Platforms such as Telegram and WhatsApp have become common tools for creating false political narratives, as the platforms host private self-moderated communities. Narratives created on these platforms were used by Bolsonaro supporters in 2018 and 2022 to both attack political opponents and delegitimize unfavorable election results (Cryst et al., 2021). A lack of legislation prohibiting this behavior allows for political elites to create distrust in voting procedures, the judiciary, and in journalists who are not overly critical of Brazil's elections.

The greatest threat of foreign interference on Brazilian internet platforms comes in the form of bots (Arnaudo, 2017). However, bots have been bought and weaponized by domestic actors, such as 2014 presidential candidates Aécio Neves and incumbent Dilma Rousseff. The bots spammed hashtags promoting their designated political candidates and attacked their opponents on Twitter, Facebook, and WhatsApp. Bot activity muddied online discourse surrounding the race by reaching millions of users and spreading contentious narratives about either candidate. It has since then become commonplace for domestic political actors to use bots leading up to elections, aimed at flooding social media channels with seemingly real users posting criticisms of politicians and election systems (Allen, 2018). After Rousseff was reelected, her usage of bots came to light sparking public outrage, and in part led to her impeachment in 2016. Botting by Neves's and Rousseff's campaigns highlighted that those not in power are able to utilize false accounts to create misleading narratives to their advantage. Current legislation does not yet outlaw creation of false accounts, leaving public discourse vulnerable to future bot attacks.

A reaction to the Marco Civil da Internet is policies drafted by left-leaning reformist politicians that punish companies for allowing content that may be false or misleading in a political or medical context. One of the more notable examples of this type of proposal is the Brazilian Law of Freedom, Responsibility and Transparency on the Internet, otherwise known as 'the Fake News Bill', which was drafted during President Bolsonaro's tenure between 2019 and 2022 but is stalled today. The Fake News Bill would require platforms to disclose their internal takedown procedures, as well as monitor and enforce takedowns on violations outlined by a Council for Transparency and Responsibility on the Internet. Noncompliance would subject platforms to a warning and a fine of up to 10% of a company's annual income (Draft Bill No. 2.630/2020). Left-leaning politicians felt that President Bolsonaro and his allies

were exploiting Brazil's broad freedom of expression protections by posting factually incorrect stories online, which the Fake News Bill aims to curtail. The general criticism of laws that require platforms to takedown speech deemed false is that it limits free speech, and additionally are said to overreach by tying users' federal identification to their online activity (Lubianco, 2022). Of the Fake News Bill, Facebook claimed it would "hinder [their] ability to limit abuse on our platforms," due to Facebook's desire to keep internal procedures used to identify and remove content private (The Times, 2021). By allowing third parties to view how platforms combat harmful content, they argue that those who wish to post harmful content will learn to circumnavigate those procedures. While the bill does detail useful provisions such as creating legal definitions for social media terminology and suppressing advertising on platforms that allow promotion of violence, it tries to accomplish too many policy goals in one piece of legislation, jeopardizing its chances at passing Congress. Without focused legislation on what is deemed misleading content and, therefore, allowable online, users will continue to post misinformation on social platforms. A lack of succinct provisions detailing misinformation continues the online climate that leaves misleading content up long enough for harmfully false narratives to spread, like that of COVID-19 misinformation.

By allowing platforms to take down content at their own discretion, some politicians argue that their free speech rights are not being upheld when their posts are removed. In September 2021, Bolsonaro issued a provisional decree that would limit platform's safe harbor allowance of content takedowns, aiming to protect political free speech and limit removals to only incitement of violence. Bolsonaro likely aimed to remove platforms' liability shields that allowed them to take down Bolsonaro's posts containing false claims about COVID vaccines (Biller, 2021). Opponents of Bolsonaro and proponents of Brazilian election integrity claim this was to allow him to continue to claim future elections are conducted unfairly (Perrigo, 2021). The decree was struck down in the Senate, which argued that it was too sweeping a change without adequate deliberation and would completely alter the framework of the Marco Civil da Internet (de Perdigão Lana et al., 2022).

Brazil's lack of disinformation and misinformation definitions leaves court rulings up to broad criminal codes that were written in a time before the current challenges of the online social landscape. Before a judge can issue orders for content removal, disinformation remains online and gains traction, reaching mainstream discourse and altering popular narratives. On the other hand, efforts to establish a government structure which counters disinformation campaigns have begun in Brazil's Judiciary. A new committee in the judiciary will support research into misinformation

to assist the drafting of legislation and create counter-narratives of disinformation and misinformation (Mega, 2022). Allocating more resources to the government will assist platforms in identifying content they may wish to remove in a timelier manner and help inform the public on related misleading narratives in lieu of related legislation.

Disputable Content: Hate Speech

The Brazilian government has legislated against certain aspects of hate speech including race, religion, and xenophobia in the late 1980's and 1990's, however it has not since expanded on what constitutes hate speech and when platforms must remove it. Brazil defines protections from insults referring to race, color, ethnicity, religion, or origin in its Racism Law (Law No. 9.459/1997). This amends Law No. 7.716/1989, which details criminal punishments relating to the media publication of discrimination or incitement of prejudice. The Racism Law can be used to jail violators from two to five years in addition to a fine. It also penalizes conveying Nazi imagery to disseminate the ideology, applying the same fine as above. Brazil deals with a high volume of hate speech complaints despite these statutes. SaferNet Brasil, a civil organization created to fight for human rights on the internet, has received more than 2.5 million hate crime complaints since 2006 (Cardoso, 2022). The University of Sao Paulo research additionally states there is not a specific criminal act of using hate speech online, only articles that address racial prejudice and violence against women (Law No. 8.081/1990). Therefore, it is up to the individuals who experience hate speech online to choose whether to pursue civil litigation as a recourse in defense of "intimacy, honor and image".

Furthermore, this legislation does not account for the evolution of hate speech, which expands to include derogatory or threatening language directed towards individuals based on their sexual or gender identity. Such online hate speech has physical consequences such as increasing murders of LGBTQIA individuals, with Brazil becoming the 'LGBT Murder Capital of the World" (Strobl, 2017). Leaving individual hate speech takedowns to platform discretion and court orders creates room for harmful content to spread and affect users. The lack of specific government guidelines highlights the gap in legislation targeting hate speech against those of marginalized identities (racial, gender, sexuality, and religion) online, beyond the country's criminal statutes.

Disputable Content: Extremism

Online extremism in Brazil has recently taken the form of terrorism and calls for political violence. The Brazilian Anti-Terrorism Law of 2016 defines terrorism as acts of prejudice based on race, color, ethnicity, or religion enacted to cause public harm or general terror. Brazil aligns its foreign terror designations of organized groups who enact violence against the public, such as ISIS and al-Qaeda, with the United States' (Counter Extremism Project, 2022). The U.S. Department of State's Antiterrorism Assistance Program provided training to Brazilian law enforcement in 2015, in part to prepare Brazil for potential attacks during the 2016 Summer Olympics. Around that time, Brazil was threatened by pro-ISIS operatives on Twitter, and ISIS propaganda was translated into Portuguese and disseminated on Telegram (Counter Extremism Project, 2022). Between 2016 and 2018, almost a dozen individuals were arrested and sentenced for recruiting and organizing online despite a lack of specific legislation detailing criminal online terror-related activity, displaying the broad application of Brazil's Anti-Terrorism Law in lieu of internet-specific anti-terror legislation. In this case, existing legislation was able to prevent potential terror attacks from occurring, but how the internet is used to incite terror and other extremism, like violent political acts, has yet to be further defined in legislation.

Brazil's 2016 Anti-Terrorism Law provides language that does not specify liability for content intermediaries; however, it is strong against actors who plot terror acts and recruit online because of Article III's language criminalizing promotion of terror organizations personally or through an intermediary (Law No. 13.260/2016). The penalty for violating Article III is imprisonment from three to five years and a fine. The language also includes recruiting for, organizing, or providing assistance to terror plots, which can be done through online messaging apps like when eight men were arrested for sharing bomb-making videos in 2017 (BBC, 2017). Between the Anti-Terrorism law and the Marco Civil, the lack of specific takedown or compliance guidelines outlining how platforms should identify and takedown terror-related online activity means potential terror-related content claims are subject to investigation by Brazil's Federal Police Force and then handled by a judge at the request of the Public Prosecutor's Office or Police Chief (Article XI & XII of Law No. 12.260/2016). Due to the importance of carefully carrying out a terror investigation, it can take time for federal investigators to collect evidence and present it to a prosecutor or judge before content is ordered to be removed via court order.

There is the possibility that the Anti-Terrorism law could be weaponized against domestic political extremists who do not hold power in the government, such

as in the case of the recent January 8th attack on Brazilian Congress instigated by Bolsonaro supporters. An indictment against 39 of the attack's perpetrators claimed that the Anti-Terror law could not be used against them, despite prior speculation that it may (FitzGerald, 2023). Prosecutors are further attempting to tie Bolsonaro's recent online posts denying the results of the election to his supporters' violent actions (Alberti, 2023). The inability to charge January 8th rioters with terrorist acts, however, may show that Brazil is unlikely to tie domestic political extremism to terrorism through legislation. In that case, online incitement of violence in a political context, such as Bolsonaro's, would not be subject to the Anti-Terrorism Law. Skirting the Law's provisions against incitement of violence makes it more challenging to act against extreme political speech within Brazil's Marco Civil framework, because platforms are not required to take down such content.

Strengths and Shortcomings

The passage of Marco Civil da Internet in 2014 has a profound influence on the digital world, dictating how one of the largest nations in the world navigates an ever-changing social landscape. The legislation succeeds in the following areas:

- It promotes democratic participation in the drafting process of liability-defining internet legislation.
- It provides specific language on what data is protected for users, procedures for requesting data by the government, and procedures for the government to pursue content takedowns.
- It serves as a foundation for future legislation to be drafted in response to further modern challenges that Brazil may experience in the coming years.

Brazil has also taken steps to create government structures centered around digital rights that support the drafting of content moderation legislation. Despite its successes, it falls short in a few areas:

- It does not create a detailed framework of what disinformation/misinformation is or when it becomes harmful and must be taken down.
- It fails to outline platform liability for allowing hate speech to propagate.
- It does not expand on the 2016 Anti-Terrorism law, leaving broad authority for state officials to act on what they deem to be online terrorist activity.
- It lacks language addressing the usage of fake accounts used to mislead.
- It does not explore the role of constitutionally prohibited anonymity in platform creation and usage.

Australia

Despite being one of the world's geographically largest countries and prominent democracies, Australia does not explicitly protect the freedom of expression in their Constitution (Freedom, 2023). However, Australia has adopted the International Covenant on Civil and Political Rights (ICCPR), which includes a section on protecting the freedom of expression (Freedom, 2023). The implied guarantee of freedom of expression through the ICCPR plus the lack of legislation that directly impedes on it means that upholding freedom of expression in Australia is not an issue. This relates to social media content since freedom of expression automatically allows for all types of content to be posted unless there are explicit prohibitions against a specific type, such as child exploitation imagery.

Australia's recent efforts in regulating social media aligns with an overall increased usage and presence of social media in Australian society (Hughes, 2023). As of February 2022, approximately 83% of Australian citizens utilized social media, compared to 58% in 2015 (Hughes, 2023). Facebook is the most popular media platform in the country with approximately 28.4 million users, followed by Instagram with 26.4 million users (Hughes, 2023). The government's attempts to increase regulation over social media content and overall stance in regard to content moderation on social media appears to focus on protecting online users from harmful content.

Australia began addressing the government's role in content moderation with the Broadcasting Services Act of 1992, which started the practice of content moderation for online services in Australia (Department, 1992). The Broadcasting Services Act set a precedent for future legislation by giving the government authority over online spaces and establishing standards over what is permitted on them. The main purpose of the Broadcasting Services Act was to protect Australians consuming broadcasted media and utilizing broadcasting services, particularly children, from encountering harmful or disturbing content (Department, 1992). Specifically, the Online Content Scheme section within this legislation establishes the government's ability to regulate or restrict harmful content (Department, 1992). While this policy primarily addresses moderation of explicitly illegal content such as child exploitation images or terrorism, it plays an important role by setting the precedent for the government to limit, restrict, and penalize other types of content. The foundation established by the Broadcasting Services Act could also allow for future legislation relating to non-illicit, but still potentially harmful content, or disputable content, to be enacted. With regards to Australia's role in moderating disputable content, their strategies are still

evolving, but there are proposed or passed pieces of legislation that are applicable to different areas of disputable content.

Disputable Content: Disinformation and Misinformation

Australia has not currently passed legislation regarding disinformation or misinformation on social media. However, as of January 2023, the federal government announced the introduction of a new untitled piece of legislation that plans to address both areas. These changes hope to decrease the amount of disinformation and misinformation on major online platforms (Samios, 2023). The Communications Minister stated this proposed act would give the Australian Communications and Media Authority (ACMA) the jurisdiction to request the removal of content spreading disinformation and misinformation on any online platform if companies do not voluntarily remove it first (Samios, 2023).

The driving force behind passing this act relates to recent ACMA studies, which uncovered a significant prevalence of misinformation in online spaces. According to their research, approximately 80% of Australian adults encountered misinformation regarding COVID-19, and 76% of adults thought online platforms should do more to reduce misinformation ("Australia to", 2022). These findings caused great concern amongst government officials and prompted legislative action to hold social media companies accountable for their role in spreading misinformation. The proposed legislation was also inspired by the Digital Industry Group Inc (DIGI), an Australian non-profit organization focused on digital advocacy, and their tightening of the voluntary Australian Code of Practice on Disinformation and Misinformation in December 2022 ("Australia's", 2023). Members of DIGI include major platforms such as Google, Apple, Meta, Twitter and TikTok, and these companies contributed to the voluntary code of practice on how to handle misinformation ("Australia's", 2023). The similar expectations between the existing DIGI code regarding disinformation and misinformation and the proposed legislation would hopefully minimize pushback from social media companies.

If the proposed disinformation and misinformation legislation were to pass, the role of the Australian government in combating misinformation would be expanded beyond content removal abilities. One proposed measure allows the ACMA to legally request social media companies' data and information on how they handle complaints and removal of content (Samios, 2023). The proposed legislation would further integrate the government's role with company content moderation policies and work towards more collaboration and transparency between the two entities.

Another proposed measure would allow the ACMA to register, create, and enforce new codes or industry standards if companies' own efforts to regulate disinformation and misinformation are inadequate (Samios, 2023). The proposed legislation would keep social media companies accountable and increase the government's involvement with future content moderation endeavors. One potential example of a new standard within the proposed legislation is requesting the company to create more resources or tools that help users identify and report disinformation and misinformation content (Samios, 2023). However, since this legislation has not officially passed, the effectiveness of enforcing new codes on social media companies is unknown.

The untitled proposed act regarding the removal of content involving disinformation and misinformation is projected to pass by the end of the year (Samios, 2023). This legislation would revolutionize Australia's ability to moderate content on the basis of disinformation and misinformation because there is no existing legislation that specifically addresses these issues. It would set a new precedent for the government's role in removing this content on social media and would give the ACMA tangible steps towards combating it. However, the true effect, outcomes, and application of the proposed legislation is still yet to be determined and it is unknown on whether it would impact the presence of disinformation and misinformation on social media within Australia.

Disputable Content: Hate Speech

Australia does not currently have legislation that addresses the challenges of defining and moderating hate speech on social media. However, the Racial Discrimination Act of 1975 is relevant to removing online content involving discriminatory language due to a federal court case (Galexia, 2002). There are also voluntary agreements regarding online safety which are applicable to hate speech, but the government lacks the ability to actually enforce the guidelines.

While the Racial Discrimination Act does not explicitly address the ability of the government to remove content on social media or define hate speech, it does ensure that racial discrimination is not perpetuated or permitted in the country. Specific language from this act includes protection from discrimination on the basis of race, ethnicity, or place of origin (Racial, 1975). In terms of the act's jurisdiction over content, it covers any written, verbal, or visual material that is made available to the public (Racial, 1975). The general nature of the act's language allows it to be interpreted and applied to a variety of scenarios, including online environments. The

Australian Human Rights Commission (AHRC) specifically references utilizing the Racial Discrimination Act's complaint process as one way to resolve incidents of racism online (Cyber, 2011).

The enforcement of the Racial Discrimination Act is the responsibility of the Australian Human Rights Commission and the Australian court system. This legislation operates on a complaint-based system where individuals who feel targeted by racially discriminating language may file with the AHRC (Racial, 1975). If the complaint is not resolved between the two parties, the complaint progresses into the court system (Racial, 1975). The court decides the outcome and compensation, including monetary funds, for the complainant.

Regarding the practical application of this legislation in an online context, the 2002 case of *Jones V. Toben* set the precedent. In this federal court case, it was determined that a website denying Holocaust and targeting Jewish people was unlawful (Galexia, 2002). The owner of the website was ordered to remove it and was found responsible for perpetuating harm to Jewish Australians and promoting offensive and disrespectful content. It was the first time this policy was applied to online material (Galexia, 2002). This court case demonstrates the plausibility of applying the Racial Discrimination Act to an online content, even though it was not on a social media platform, and the act's broad parameters allow for flexibility on what is considered discriminatory content. While this process allows for a case-by-case basis decision on racially discriminating language, it lacks a standardized set of outcomes in terms of penalties.

The need for specific hate speech legislation aligns with research from the eSafety Commissioner that found 14% of adults ages 18-65 were targeted by hate speech in 2019 (Online Hate, 2019). The same study also found that people identifying as LGBTQ, Aboriginal, or Torres Strait Islander were more likely to encounter hate speech online (Online Hate, 2019). Besides conducting research on hate speech, the eSafety Commissioner also created a set of voluntary guidelines through the Online Safety Charter (Online, 2019). This charter helps social media companies address online safety issues and creates responsibilities for companies. (Online, 2019). While the Online Safety Charter created general parameters for social media content and does not specifically focus on hate speech, it remains an important agreement because it requires social media companies to adhere to certain safety principles and could be applied to cases of hate speech. In addition to the Online Safety Charter, the AHRC created a report on protecting human rights in online spaces that includes guidelines and proposals for promoting online safety and reducing hate speech (Background, 2013). Like the eSafety Commissioner's guidelines, this agreement is not

legally binding so cannot be enforced. While it is evident that guidelines around reducing hate speech online exist in Australia, their effectiveness is limited because they are voluntarily followed and are not official legislation.

While the Racial Discrimination Act can address hate speech, it does fall short because it is not clear how the government would hold social media companies liable for hate speech on their platform. When surveying Australian adults, the eSafety Commissioner found a majority supported more stringent measures for hate speech, including passing legislation and increasing social media company liability (Online, 2019). The Racial Discrimination Act is designed for complaints filed against individuals versus companies and it is unclear how a company's roles would be adapted to this process if a complaint was filed against them. Part of the issue is the company's role as a third party in spreading hate speech since this is distinct from creating content.

Furthermore, the Racial Discrimination Act does not specifically define what hate speech is, which could cause confusion over what language is considered lawful on platforms—ambiguity is something that many countries are struggling with in terms of content moderation, and Australia will certainly need to consider it when crafting this legislation. While the eSafety Commissioner has a common definition of hate speech, this is not reflected in any current legislation (Online, 2023). Overlap between the content addressed in the Racial Discrimination Act and hate speech likely exists, but without a clear definition of hate speech, it's difficult to apply the act overarchingly. Due to *Jones V. Toben,* it is reasonable to assume that the Australian government may continue to utilize the Racial Discrimination Act to remove and restrict content online, including on social media platforms. However, future limitations or uncertainty regarding what content is regulated under this act may arise because it does not explicitly address a company's role or define hate speech. The passing of voluntary agreements on hate speech created by the eSafety Commissioner and the AHRC into law is another potential method of decreasing its presence on social media and would formalize these current agreements.

Disputable Content: Extremism

Australia has legislation regarding online content and extremism, but it needs to be expanded in order to regulate disputable content. This piece of current legislation, the Criminal Code Amendment (Sharing of Abhorrent Violent Material Act) was created to address extremely violent online content in response to the Christchurch Massacre (Christchurch, 2020). The Christchurch Massacre occurred on March 15th of 2019, leaving 51 dead and 40 injured. During the massacre, 17 minutes

of the mass shooting at two mosques in New Zealand were broadcast live on Facebook (Christchurch, 2020).

The Criminal Code Amendment (Sharing of Abhorrent Violent Material Act) made violent online conduct a criminal offense (Abhorrent, 2021). The amendment covers material in the form of images, audio, or video and the abhorrent illegal violent material includes acts of murder, terrorism, torture, rape, and kidnapping ("Australia can", 2019). This amendment increased the obligations of individuals and online service providers by requiring them to report violent content to the Australian Federal Police within a reasonable time (Abhorrent, 2021). For reporting, it is acknowledged situations differ so determining a reasonable time frame is contingent on the type and volume of the material, and how many users reported it to the company (Abhorrent, 2021). The obligation to report content is applicable to both individuals and companies who encounter it, but the removal of abhorrent content falls solely on the content providers (Abhorrent, 2021). It further specified that violent content includes anything occurring in Australia, accessible through services found in Australia, or is hosted through services found in Australia even if the company is located in another country (Abhorrent, 2021).

The eSafety Commissioner of Australia is the enforcement agency for this law and handles the reporting and removal process for violent content. The penalties established differ for individuals or companies. For an individual who fails to remove content, a fine of up to 2.22 million AUD and up to three years of prison time is possible (Abhorrent, 2021). For companies, a fine of up to 11.1 million AUD or 10% of annual company turnover may be imposed, whichever monetary amount is greater for that company (Abhorrent, 2021). However, the amendment acknowledges certain situations where individuals or companies were unaware of the violent content and does not penalize them. It also does not hold companies liable for violent content published that relates to law enforcement purposes, news and current affairs, or public policy and research reasons (Abhorrent, 2021). The amendment only penalizes those who were aware of the content, but failed to remove or report it (Abhorrent, 2021). However, it is difficult to prove whether the company was aware or not and the eSafety Commissioner's exact methodology and process for determining awareness is not clear.

The 2019 amendment also falls short because it does not address liability for certain extremist online content, such as inciting violence or recruitment into terrorist organizations. The language of the amendment is solely limited to the illicit content listed in the law. This creates gaps regarding the liability of individuals and companies for the violent acts not listed and limits the jurisdiction of the eSafety Commissioner

for penalizing this disputable content. There is not a current method for regulating and reporting disputable content related to extremism, but the creation of the 2019 amendment demonstrates that the government is willing and able to place liability on companies for illicit acts. While this legislation was designed to combat violent and extreme content, it may face challenges in creating accountability for companies and their role in the promotion of disputable content. It has, however, created a foundation that future legislation may be able to expand upon should Australia decide to do so.

Strengths and Shortcomings

Australia has proposed and passed content moderation related legislation that seeks to address disputable content. This is demonstrated by proposed legislation to reduce the presence of disinformation and misinformation, Australia's willingness to amend existing legislation, and placing liability on platforms for certain types of content. These strengths have created a solid basis for the government's content moderation abilities and placing responsibility on the companies as needed. Their strengths are as follows:

- The proposal of legislation specifically designed for combating misinformation and disinformation highlights an understanding of disputable content and the need for government involvement in minimizing it. This indicates a new interest in creating legislation to address recent trends in disputable online content.
- The amendment of the Criminal Code regarding online content showcases a willingness to adapt current laws based on current needs and to help bridge existing gaps. This indicates Australia's understanding of current laws' shortcomings along with the ability to recognize these areas of uncertainty which prompts them to develop solutions.
- Proposed and passed legislation created penalties for companies not in compliance. These accountability measures, whether new industry codes or fines, demonstrate a commitment to holding companies accountable and placing responsibility for certain content moderation on them.

While significant progress regarding content moderation in Australia has occurred, there are definite areas for improvement to bridge gaps and clarify uncertainties in current legislation relating to clarity of definitions and increasing transparency around current moderation practices. Future legislation or amendments involving these issues would strengthen the overall approach to content moderation

by increasing the jurisdiction of the government and refining expectations for companies. Despite its successes, it falls short in a few areas:

- The lack of clear and specific definitions and legislation may increase the government's ability to address disputable content. Clarification is helpful for enforcement agencies, such as the ACMA or the court system, in order to give them a concise and singular reference point for what online content is not permitted.
- Gaps in legislation exist regarding the liability of companies for extremist content that are not listed in the Criminal Code amendment. Penalties are only issued in clearly illicit situations which allows for disputable content to remain online without creating consequences for companies.
- The process for determining whether companies were aware of content and assessing their role in promoting it could become more transparent since the discretion of enforcement entities is relied on for decisions. Since companies are expected to voluntarily remove it, this makes discerning whether they were aware or not difficult to prove.

Bibliography

Executive Summary and Findings Sources

Barrett, Paul. *A Call to End Outsourcing - static1.Squarespace.com*. NYU Stern Center for Business and Human Rights, June 2020, https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU%20Content%20Moderation%20Report_June%208%202020.pdf

Capoot, Ashley. *Meta Is Rolling out a New Paid Verification Subscription Service for Instagram and Facebook Users*. CNBC, 19 Feb. 2023, https://www.cnbc.com/2023/02/19/meta-is-rolling-out-a-new-paid-verification-subscription-service-for-instagram-and-facebook-users.html.

Corrons, Luis. *Twitter Blue Means Bad Things for the Platform's Security*, Nov. 2022, https://blog.avast.com/twitter-blue-security#:~:text=The%20potential%20for%20malicious%20threat,politician%20is%20impersonated%20on%20Twitter.

Fioretti, Julia. *Social Media Giants Step up Joint Fight against Extremist Content*. Thomson Reuters, 26 June 2017, https://www.reuters.com/article/us-internet-extremism/social-media-giants-step-up-joint-fight-against-extremist-content-idUSKBN19H20A.

Gonzalez v. Google LLC. Oyez, 2023, https://www.oyez.org/cases/2022/21-1333.

Mac, Ryan, et al. *A Verifiable Mess: Twitter Users Create Havoc by Impersonating Brands*. The New York Times, 11 Nov. 2022, https://www.nytimes.com/2022/11/11/technology/twitter-blue-fake-accounts.html.

McCain, Abby. *Zippia 40+ Trending Facebook Statistics 2023 Revenue Usage Demographics Statistics For Marketers Comments*, Zippia, Dec. 2022, https://www.zippia.com/advice/facebook-statistics/.

Miller, Robert T. *What Is a Compelling Governmental Interest?* University of Iowa College of Law, 26 Mar. 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3149162.

Perez, Sarah. *Tiktok to Flag and Downrank 'Unsubstantiated' Claims Fact Checkers Can't Verify, TechCrunch*, 3 Feb. 2021, https://techcrunch.com/2021/02/03/tiktok-to-flag-and-downrank-unsubstantiated-claims-fact-checkers-cant-verify/.

Samples, John. *Why the Government Should Not Regulate Content Moderation of Social Media*, CATO Insitute, 2019, https://www.cato.org/policy-analysis/why-government-should-not-regulate-content-moderation-social-media#.

Singh, S., & Doty, L. (December 9, 2021). The Transparency Report Tracking Tool: How internet platforms are reporting on the enforcement of their content rules. New America. Retrieved March 2, 2023, from https://www.newamerica.org/oti/reports/transparency-report-tracking-tool/

Walsh, Shelley. *The Top 10 Social Media Sites & Platforms 2022." Search Engine Journal*, 12 Aug. 2022, https://www.searchenginejournal.com/social-media/biggest-social-media-sites/#close.

Disinformation, Misinformation, Hate Speech, & Extremism Sources

"Alt Right: A Primer on the New White Supremacy." *Www.adl.org*, 10 Feb. 2016, www.adl.org/resources/backgrounder/alt-right-primer-new-white-supremacy.

Atske, S. (2021, February 22). *Misinformation and Competing Views of Reality Abounded Throughout 2020*. Pew Research Center's Journalism Project. Retrieved from https://www.pewresearch.org/journalism/2021/02/22/misinformation-and-competing-views-of-reality-abounded-throughout-2020/

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., & Volfovsky, A. (2019). Assessing the Russian internet research agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. Proceedings of the National Academy of Sciences, 117(1), 243-250. https://doi.org/10.1073/pnas.1906420116

BBC. (2020, April 24). *Coronavirus: Outcry after Trump suggests injecting disinfectant as treatment*. BBC News. Retrieved from https://www.bbc.com/news/world-us-canada-52407177

Benner, K., & AFraniere, S. L. (2020, March 16). Justice Dept. Moves to Drop Charges Against Russian Firms Filed by Mueller. The New York Times. https://www.nytimes.com/2020/03/16/us/politics/concord-case-russian-interference.html

Boseley, S. (2020, May 22). *Hydroxychloroquine: Trump's COVID-19 'cure' increases deaths, global study finds*. The Guardian. Retrieved from https://www.theguardian.com/science/2020/may/22/hydroxychloroquine-trumps-COVID-19-cure-increases-deaths-global-study-finds

Bradshaw, S., & Howard, P. (2017). Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. In Computational Propaganda Research Project (pp. 1–37). Oxford Internet Institute.

Carter, Camden. (2022, February 18). *Instagram is letting accounts promoting hate speech go unchecked*. Media Matters for America. Retrieved February 25, 2023, from https://www.mediamatters.org/facebook/instagram-letting-accounts-promoting-hate-speech-go-unchecked

Centers for Disease Control and Prevention. (2020, April 23). *Cleaning and disinfectant chemical exposures and temporal associations with COVID-19 - National Poison Data System, United States, January 1, 2020–March 31, 2020*. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/mmwr/volumes/69/wr/mm6916e1.htm

Chou, W.-Y. S., Gaysynsky, A., & Cappella, J. (2021, May). *Where we go from here: Health misinformation on social media*. Pan American Journal of Public Health. Retrieved from https://www.paho.org/journal/en/articles/where-we-go-here-health-misinformation-social-media

Collins, K. (2018, May 14). See which Facebook Ads Russians Targeted to People Like You. The New York Times. https://www.nytimes.com/interactive/2018/05/14/technology/facebook-ads-congress.html

Daniels, Jesse. (2018). The Algorithmic Rise of the "Alt-Right." *Contexts*, *17*(1), 60–65. https://doi.org/10.1177/1536504218766547

Devitt, J. (n.d.). Exposure to Russian Twitter campaigns in 2016 presidential race highly concentrated, largely limited to strongly partisan Republicans. NYU Web Communications. https://www.nyu.edu/about/news-publications/news/2023/january/exposure-to-russian-twitter-campaigns-in-2016-presidential-race-.html

Díaz, Ángel and Hecht-Felella, Laura (2021) *Double Standards in Social Media Content Moderation.* Brennan Center for Justice. August 4th, 2021. https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J., & Ohnson, B. J. (2019, October). The Tactics & Tropes of the Internet Research Agency. University of Nebraska - Lincoln. https://digitalcommons.unl.edu/senatedocs/2/

Domonoske, Camila. "I Guess We Need to Talk about Pepe the Frog." *NPR*, 28 Sept. 2016, www.npr.org/sections/thetwo-way/2016/09/28/495760153/i-guess-we-need-to-talk-about-pepe-the-frog.

Evanega, S., Lynas, M., Adams, J., & Smolenyak, K. (2021, January 1). *Coronavirus Misinformation: Quantifying Sources and Themes in the COVID-19 'Infodemic': Semantic Scholar*. Coronavirus misinformation: quantifying sources and themes in the COVID-19 'infodemic' | Semantic Scholar. Retrieved from https://www.semanticscholar.org/paper/Coronavirus-misinformation%3A-quantifying-sources-and-Evanega-Lynas/b53c87652de95de0c34eaca103737461b1ee89e7

Fake news alert: Taiwan fights disinformation as COVID surges. (2021, May 28). Al Jazeera. https://www.aljazeera.com/news/2021/5/28/disinformation-goes-viral-as-taiwan-battles-new-COVID-surge

Farokhmanesh, M. (2018, May 25). *Facebook has an official Pepe the Frog policy*. The Verge. https://www.theverge.com/2018/5/25/17394612/facebook-pepe-frog-ban-meme-anti-semitic-hate

Foreign, Commonwealth & Development Office, The RT Hon Elizabeth Truss MP, & The RT Hon Nadine Dorries MP. (2022, May 1). UK exposes sick Russian troll factory plaguing social media with Kremlin propaganda. GOV.UK. https://www.gov.uk/government/news/uk-exposes-sick-russian-troll-factory-plaguing-social-media-with-kremlin-propaganda

Frenkel, S., Alba, D., & Zhong, R. (2020, March 8). Surge of Virus Misinformation Stumps Facebook and Twitter. The New York Times. https://www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html

Gadde, V., & Beykpour, K. (2020, November 12). *An update on our work around the 2020 US elections*. Twitter. Retrieved from https://blog.twitter.com/en_us/topics/company/2020/2020-election-update

Geldsetzer, P. (2020, July 21). *Knowledge and perceptions of COVID-19 among the general public in the United States and the United Kingdom: A cross-sectional online survey*. U.S. National Library of Medicine. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7086377/

Global Internet Forum to Counter Terrorism (GIFCT), *Content-Sharing Algorithms, Processes, and Positive Interventions Working Group Part 1: Content-Sharing Algorithms & Processes*. (2021). https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPI1-

2021.pdf#:~:text=While%20users%20share%20content%20and%20algorithms%20organize%20information%2C

Harold, S., Beauchamp-Mustafaga, N., & Hornung, J. (2021). Chinese disinformation efforts on social media. RAND Corporation. https://doi.org/10.7249/rr4373.3

"Hate Speech and Hate Crime", American Library Association, December 12, 2017. http://www.ala.org/advocacy/intfreedom/hate (Accessed January 18, 2023) Document ID: aa35c1c7-f3aa-4b07-964f-30dcf85a503c

Heilweil, R., & Ghaffary, S. (2021, January 8). *How trump's internet built and broadcast the capitol insurrection*. Vox. Retrieved from https://www.vox.com/recode/22221285/trump-online-capitol-riot-far-right-parler-twitter-facebook

Hossain, M. S. (2015). Social Media and Terrorism: Threats and Challenges to the Modern Era. *South Asian Survey*, *22*(2), 136–155. https://doi.org/10.1177/0971523117753280

Howard, P., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2019). The IRA, social media and political polarization in the United States, 2012–2018. University of Nebraska - Lincoln, 3(1), 92-92. https://doi.org/10.1515/sirius-2019-1017

Ipsos. (2022, January 3). *Seven in ten Americans say the country is in crisis, at risk of failing*. Ipsos. Retrieved from https://www.ipsos.com/en-us/seven-ten-americans-say-country-crisis-risk-failing

Jakubowicz, A. 2017. Alt_Right White Lite: trolling, hate speech and cyber racism on social media. Cosmopolitan Civil Societies: an Interdisciplinary Journal. 9(3), 41-60. http://dx.doi.org/10.5130/ccs.v9i3.5655

Jones, Z. C., & Bey, J. (2021, January 13). *Twitter permanently bans trump*. CBS News. Retrieved from https://www.cbsnews.com/news/trump-twitter-account-suspended-permanently/

Kaminski, Margot E., Incitement to Riot in the Age of Flash Mobs (February 21, 2013). University of Cincinnati Law Review, Vol. 81, 1 2013, https://ssrn.com/abstract=2222160

Kalathil, S. (2020). The Evolution of Authoritarian Digital Influence: Grappling with the New Normal. *PRISM*, *9*(1), 32–51. https://www.jstor.org/stable/26940158

Kim, Y. M. (2020, March 5). New evidence shows how Russia's election interference has gotten more brazen. Brennan Center for Justice.

Kim, Y.M. (2018). Uncover: Strategies and Tactics of Russian Interference in US Elections Russian Groups Interfered in Elections with Sophisticated Digital Campaign Strategies.

Kim, Y. M., Hsu, J., Neiman, D., Kou, C., Bankston, L., Kim, S. Y., Heinrich, R., Baragwanath, R., & Raskutti, G. (2018). The stealth media? Groups and targets behind divisive issue campaigns on Facebook. Political Communication, 35(4), 515-541. https://doi.org/10.1080/10584609.2018.1476425

Klar, R. (2022, September 16). *Social media platforms' 'flawed policies' amplify election fraud claims: Report* . The Hill. Retrieved from https://thehill.com/policy/technology/3646955-social-media-platforms-flawed-policies-amplify-election-fraud-claims-report/

Klein, A. (2017). *Fanaticism, Racism, and Rage Online: Corrupting the Digital Sphere*. Springer International Publishing AG. https://doi.org/10.1007/978-3-319-51424-6

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021, February 5). *Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA*. Nature News. Retrieved from https://www.nature.com/articles/s41562-021-01056-1

Lovelace Jr., B. (2020, May 29). *Hydroxychloroquine prescriptions surged 2,000% in March when Trump touted malaria drug*. CNBC. Retrieved from https://www.cnbc.com/2020/05/29/coronavirus-hydroxychloroquine-prescription-fills-surged-in-march-after-trump-touted-drug.html

Lührmann, A., Gastaldi, L., Grahn, S., Lindberg, S., Maxwell, L., Mechkova, V., Morgan, R., Stepanova, N., & Pillai, S. (2019). Varieties of Democracy (V-DEM) Annual Report 2019 - "Democracy Facing Global Challenges". V-Dem Institute, the Department of Political Science at University of Gothenburg, 1-76. http://hdl.handle.net/10993/43444

Ma, A. (2019, November 28). *The 17-year-old who got suspended by Tiktok after posting viral videos about oppressed Muslims accuses the company of lazy, racist assumptions about Islam and terrorism*. Business Insider. Retrieved February 25, 2023, from https://www.businessinsider.com/tiktok-uighur-teen-feroza-slams-app-anti-muslim-2019-11

Macdonald, S., Correia, S., & Watkin, A. (2019). Regulating terrorist content on social media: Automation and the rule of law. *International Journal of Law in Context, 15*(2), 183-197. doi:10.1017/S1744552319000119

Mariash, Alexis, (2022) *Addressing the Alt-Right Pipeline*. Berkeley Political Review. October, 31st, 2022, https://bpr.berkeley.edu/2022/10/31/addressing-the-alt-right-pipeline/#:~:text=The%20alt-right%20pipeline%20is%20the%20path%20people%20follow,reality%20to%20overcompensate%20for%20active%20progress%20in%20society.

Meta. (2023, February 7). *Violence and incitement*. Transparency Center. Retrieved from https://transparency.fb.com/policies/community-standards/violence-incitement/

Mourali, M., & Drake, C. (2022, February 3). *The challenge of debunking health misinformation in dynamic social media conversations: Online randomized study of public masking during COVID-19*. Journal of Medical Internet Research. Retrieved from https://www.jmir.org/2022/3/e34831/

National Institutes of Health. "Implicit Bias | SWD at NIH." *Diversity.nih.gov*, 3 June 2022, diversity.nih.gov/sociocultural-factors/implicit-bias. Accessed 30 Jan. 2023.

NPR/PBS NewsHour/Marist National Poll. (2021, November). *Trust in elections & threat to democracy, Nov 2021*. Marist Poll. Retrieved from https://maristpoll.marist.edu/polls/npr-pbs-newshour-marist-national-poll-trust-in-elections-threat-to-democracy-biden-approval-november-2021/

Offices of the United States Attorneys. (2021, December 30). *One year since the Jan. 6 attack on the Capitol*. The United States Department of Justice. Retrieved from https://www.justice.gov/usao-dc/one-year-jan-6-attack-capitol

Pearson, Nick. "Pepe the Frog Meme Now Officially a Hate Symbol." *Www.9news.com.au*, 28 Sept. 2016, www.9news.com.au/world/pepe-the-frog-meme-now-officially-a-hate-symbol/5a1f644c-7d46-47ce-bb43-99181619c9a3.

Quinnipiac University. (2022, January 12). *Political instability not U.S. adversaries, seen as bigger threat, Quinnipiac University National Poll finds; nearly 6 in 10 think nation's democracy is in danger of collapse: Quinnipiac University poll*. Political Instability Not U.S. Adversaries, Seen As Bigger Threat, Quinnipiac University National Poll Finds; Nearly 6 In 10 Think Nation's Democracy Is In Danger Of Collapse | Quinnipiac University Poll. Retrieved from https://poll.qu.edu/poll-release?releaseid=3831

Quirk, S. (2021). Lawfare in the disinformation age: chinese interference in taiwan's 2020 elections. *Harvard International Law Journal, 62(2),* 525-568.

Ratliff, A. (2021, November 17). *60% of president Donald Trump's post-election tweets sought to undermine legitimacy of presidential race*. Issue One. Retrieved from https://issueone.org/articles/60-of-president-donald-trumps-post-election-tweets-sought-to-undermine-legitimacy-of-presidential-race/

Rubin, O., Mallin, A., & Steakin, W. (2022, January 4). *By the numbers: How the Jan. 6 investigation is shaping up 1 year later*. ABC News. Retrieved from https://abcnews.go.com/US/numbers-jan-investigation-shaping-year/story?id=82057743

Ruckenstein, M., & Turunen, L. L. M. (2020). Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society*, *22*(6), 1026–1042. https://doi.org/10.1177/1461444819875990

Sanderson, Z., Brown, M. A., Bonneau, R., Nagler, J., & Tucker, J. A. (2021, August 24). *Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform: HKS Misinformation Review*. Misinformation Review. Retrieved from https://misinforeview.hks.harvard.edu/article/twitter-flagged-donald-trumps-tweets-with-election-misinformation-they-continued-to-spread-both-on-and-off-the-platform/

"Social Media and Terrorism a Dangerous Mechanism." *Global Counter Terrorism Council*, GCTC, 1 Dec. 2021, https://www.gctcworld.org/social-media-and-terrorism-a-dangerous-mechanism/.

Speckhard, A., & Ellenberg, M. D. (2020). ISIS in Their Own Words: Recruitment History, Motivations for Joining, Travel, Experiences in ISIS, and Disillusionment over Time – Analysis of 220 In-depth Interviews of ISIS Returnees, Defectors and Prisoners. *Journal of Strategic Security*, *13*(1), 82–127. https://www.jstor.org/stable/26907414

SQL Server Reporting Services. (2021, September 15). *Embargoed for release: Wednesday, September 15 at noon - CNN*. Cable News Network. Retrieved February 8, 2023, from http://cdn.cnn.com/cnn/2021/images/09/15/rel5e.-.elections.pdf

Stolberg, S., & Weiland, N. (2020, September 30). *Study finds 'single largest driver' of coronavirus misinformation: Trump*. The New York Times. Retrieved from https://www.nytimes.com/2020/09/30/us/politics/trump-coronavirus-misinformation.html?register=email&auth=register-email

Torok, Robyn. "Isis and the Institution of Online Terrorist Recruitment." *Middle East Institute*, 1 Jan. 2022, https://www.mei.edu/publications/isis-and-institution-online-terrorist-recruitment.

'Troll factory' spreading Russian pro-war lies online, says UK. (2022, May 1). The Guardian. https://www.theguardian.com/world/2022/may/01/troll-factory-spreading-russian-pro-war-lies-online-says-uk

Twitter. (2021, January 8). *Permanent suspension of @realDonaldTrum*. Twitter. Retrieved from https://blog.twitter.com/en_us/topics/company/2020/suspension

Wang, J. C. (2020, April 7). *Hydroxychloroquine: How an unproven drug became Trump's Coronavirus 'Miracle Cure'*. The Guardian. Retrieved from

https://www.theguardian.com/world/2020/apr/06/hydroxychloroquine-trump-coronavirus-drug

Ward, A. (2018, December 11). *Isis's social media use poses a threat to stability in the Middle East and Africa*. RAND Corporation. Retrieved February 25, 2023, from https://www.rand.org/blog/2018/12/isiss-use-of-social-media-still-poses-a-threat-to-stability.html

Weedon, J., Nuland, W., & Stamos, A. (2017). Information operations and Facebook. *Retrieved from Facebook: https://fbnewsroomus. files. wordpress. com/2017/04/facebook-and-information-operations-v1. pdf*.

Weimann, G. (2016). Terrorist Migration to the Dark Web. *Perspectives on Terrorism*, *10*(3), 40–44. http://www.jstor.org/stable/26297596

Wiggins, B. E., & Bowers, G. B. (2015). Memes as genre: A structurational analysis of the memescape. *New Media & Society*, *17*(11), 1886–1906. https://doi.org/10.1177/1461444814535194

Wolf, Z. B. (2021, May 19). *The 5 key elements of Trump's Big Lie and how it came to be | CNN politics*. CNN. Retrieved from https://www.cnn.com/2021/05/19/politics/donald-trump-big-lie-explainer/index.html

World Health Organization. (2023). *Infodemic*. World Health Organization. Retrieved from https://www.who.int/health-topics/infodemic/the-COVID-19-infodemic#tab=tab_1

Wray, C. (2017, November 30). *Keeping America Secure in the New Age of Terror*. Federal Bureau of Investigation. https://www.fbi.gov/news/testimony/keeping-america-secure-in-the-new-age-of-terror

Content Moderation & Social Media Platform Sources

About ADL's work combating cyberhate and Countering Violent Extremists Online. ADL. (2016, February 29). Retrieved February 26, 2023, from https://www.adl.org/resources/news/about-adls-work-combating-cyberhate-and-countering-violent-extremists-online

AI Content Moderators. Help center. (n.d.). Retrieved January 24, 2023, from https://help.instagram.com/423837189385631

Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan, Showkat Hassan Malik, Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques, International Journal of Information Management Data Insights, Volume 2, Issue 2, 2022

Arsht, A., & Etcovitch, D. (2018, March 2). *The human cost of online content moderation*. Harvard Journal of Law & Technology. Retrieved February 26, 2023, from https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation

Ayad, M. (2019, August 9). *Facebook and YouTube are failing to detect terrorist content in Arabic*. VICE. Retrieved February 2, 2023, from https://www.vice.com/en/article/59nmyd/facebook-and-youtube-are-failing-to-detect-terrorist-content-in-arabic

Baker-White, Emily (2022) How TikTok Has Bent It's Rules For Its Top Creators, https://www.forbes.com/sites/emilybaker-white/2022/09/20/tiktok-special-treatment-top-creators-bent-rules/?sh=22de67c8590c

Barry, Eliose (2022) These Are The Countries Where Twitter, Facebook, and TikTok are Banned. Time.

BBC. (2012, June 29). *China blocks access to Bloomberg and BusinessWeek sites*. BBC News. Retrieved February 1, 2023, from https://www.bbc.com/news/technology-18648050

BBC. (2020, February 27). YouTube 'not a public forum' with guaranteed free speech. BBC News. Retrieved February 26, 2023, from https://www.bbc.com/news/technology-51658341

Beckett, L. (2020, October 10). *Michigan terror plot: why rightwing extremists are thriving on Facebook*. The Guardian. Retrieved February 16, 2023, from https://www.theguardian.com/technology/2020/oct/09/facebook-rightwing-extremists-michigan-plot-militia-boogaloo

Beckett, L. (2020, April 30). *Armed protesters demonstrate against COVID-19 lockdown at Michigan capitol*. The Guardian. Retrieved February 16, 2023, from https://www.theguardian.com/us-news/2020/apr/30/michigan-protests-coronavirus-lockdown-armed-capitol

Berger, J., & Perez, H. (2016). *Occasional Paper The Islamic State's Diminishing Returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters*. Retrieved from https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/downloads/JMB%20Diminishing%20Returns.pdf

Biggest social media platforms 2022. Statista. (2022, July 26). Retrieved February 1, 2023, from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Blystone, D. (2022, October 22). *Instagram: What it is, its history, and how the Popular App Works*. Investopedia. Retrieved January 30, 2023, from

https://www.investopedia.com/articles/investing/102615/story-instagram-rise-1-photo0sharing-app.asp

Bond, S. (2021, March 9). *Instagram suggested posts to users. it served up COVID-19 falsehoods, study finds*. NPR. Retrieved January 31, 2023, from https://www.npr.org/2021/03/09/975032249/instagram-suggested-posts-to-users-it-served-up-COVID-19-falsehoods-study-finds

Brouwer, B. (2015, October 12). *YouTube expands global reach, boasts 85 local versions of its site*. Tubefilter. Retrieved February 17, 2023, from https://www.tubefilter.com/2015/10/12/youtube-global-expansion-85-local-versions/

Carman, A. (2021, February 10). *Instagram announces tougher consequences for hate speech in direct messages*. The Verge. Retrieved February 19, 2023, from https://www.theverge.com/2021/2/10/22276491/instagram-direct-message-hate-speech-account-disabled-policy

Carson, M. (2022, October 19). *What age verification technology does in social media*. Aspioneer. Retrieved February 26, 2023, from https://aspioneer.com/what-age-verification-technology-does-in-social-media/#:~:text=Social%20media%20platforms%20invest%20in%20age%20verification%20tools,is%20there%20to%20protect%20minors%20from%20harmful%20content.

Cascini, F., Pantovic, A., Al-Ajlouni, Y. A., Failla, G., Puleo, V., Melnyk, A., Lontano, A., & Ricciardi, W. (2022, June). *Social media and attitudes towards a COVID-19 vaccination: A systematic review of the literature*. EClinicalMedicine. Retrieved February 16, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120591/

Castella, T. de, & Brown, V. (2011, September 14). *Trolling: Who does it and why?* BBC News. Retrieved February 26, 2023, from https://www.bbc.com/news/magazine-14898564

Child Sexual Exploitation, Abuse and Nudity. Transparency Center. (n.d.). Retrieved January 24, 2023, from https://transparency.fb.com/policies/community-standards/child-sexual-exploitation-abuse-nudity/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fchild_nudity_sexual_exploitation

Coleman, K. (2019, March 30). *Instagram has a problem with hate speech and extremism, 'Atlantic' reporter says*. NPR. Retrieved January 31, 2023, from https://www.npr.org/2019/03/30/708386364/does-instagram-have-a-problem-with-hate-speech-and-extremism

Community Guidelines. Help center. (n.d.). Retrieved January 24, 2023, from https://help.instagram.com/477434105621119

Community standards enforcement. Transparency Center. (n.d.). Retrieved January 31, 2023, from https://transparency.fb.com/data/community-standards-enforcement/

Community guidelines: Facebook help center. Facebook. (n.d.). Retrieved January 18, 2023, from https://www.facebook.com/help/477434105621119

"Community Guidelines." TikTok. https://www.tiktok.com/community-guidelines?lang=en.

Congress. (n.d.). Retrieved February 2, 2023, from https://crsreports.congress.gov/product/pdf/R/R46662

Cox, J. (2018, May 14). *Leaked documents show how Instagram polices content to prevent 'PR fires'*. VICE. Retrieved February 1, 2023, from https://www.vice.com/en/article/59qpbk/leaked-instagram-content-moderation-guidelines

Criddle, C. (2021, July 15). *Instagram admits moderation mistake over racist comments*. BBC News. Retrieved February 19, 2023, from https://www.bbc.com/news/technology-57848106

D'Anastasio, C. (2020, December 9). A Christchurch report points to YouTube's radicalization trap. Wired. Retrieved February 26, 2023, from https://www.wired.com/story/christchurch-shooter-youtube-radicalization-extremism/

de Keulenaar, Emillie, João C. Magalhães, and Bharath Ganesh. 2022. "Modulating Moderation: A Genealogy of Objectionable Content on Twitter." MediArXiv. September 1. doi:10.33767/osf.io/wvp8c.

Delkic, Melina. "Leg Booty? Panoramic? Seggs? How Tiktok Is Changing Language." The New York Times. The New York Times, November 19, 2022. https://www.nytimes.com/2022/11/19/style/tiktok-avoid-moderators-words.html.

De Vynck, G. (2021, April 6). YouTube says it's getting better at taking down videos that break its rules. they still number in the millions. The Washington Post. Retrieved February 26, 2023, from https://www.washingtonpost.com/technology/2021/04/06/youtube-video-ban-metric/

Diaz, Angél; Hecht-Fellela, Laura. (2021, August 4). *Double standards in social media content moderation*. Brennan Center for Justice. Retrieved February 20, 2023,

from https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf

Dixon, S. (2022, July 27). *Meta: annual revenue and net income 2007-2021*. Statista. Retrieved January 29, 2023, from https://www.statista.com/statistics/277229/facebooks-annual-revenue-and-net-income/

DOJ. (2014). *Twitter and Violent Extremism*. Retrieved from https://cops.usdoj.gov/RIC/Publications/cops-w0741-pub.pdf

Donoghue, C. (n.d.). *Instagram and black censorship: Where did it all go so wrong?* SHOWstudio. Retrieved January 31, 2023, from https://www.showstudio.com/news/instagram-and-black-censorship-where-did-it-all-go-so-wrong

Downey, L. (2021, September 2). Google's incredible YouTube purchase 15 years later. Investopedia. Retrieved February 26, 2023, from https://www.investopedia.com/google-s-incredible-youtube-purchase-15-years-later-5200225

Doyle, Brandon. "Tiktok Statistics - Everything You Need to Know [Jan 2023 Update]." Wallaroo Media. Retrieved January 2023, from https://wallaroomedia.com/blog/social-media/tiktok-statistics/.

Espada, Mariah. "Why Are Universities Banning TikTok? What to Know." Time. Time, January 24, 2023. https://time.com/6249522/public-universities-banning-tiktok/.

Extremist content online: Instagram edition. Counter Extremism Project. (2022, November 10). Retrieved February 19, 2023, from https://www.counterextremism.com/press/extremist-content-online-instagram-edition

Facebook. (n.d.). *Facebook community standards*. Transparency Center. Retrieved February 16, 2023, from https://transparency.fb.com/policies/community-standards/

Faria, J. (2023, January 6). *Instagram Brand Value 2022*. Statista. Retrieved January 30, 2023, from https://www.statista.com/statistics/1324427/instagram-brand-value/

Ferghana.ru. (2009, December 25). *Turkmenistan: YouTube and LiveJournal are blocked*. archive.ph. Retrieved February 1, 2023, from https://archive.ph/20120707171901/http://enews.ferghananews.com/news.php?id=1516&mode=snews

Finkle, E. (2022, December 5). *Bringing age verification to Facebook Dating*. Facebook. Retrieved February 16, 2023, from https://about.fb.com/news/2022/12/facebook-dating-age-verification/#:~:text=Starting%20today%2C%20we're%20expanding,prevent%20minors%20from%20accessing%20it.

France-Presse, A. (2021, August 25). *YouTube says it has removed 1 million 'dangerous' videos on COVID*. VOA. Retrieved February 17, 2023, from https://www.voanews.com/a/silicon-valley-technology_youtube-says-it-has-removed-1-million-dangerous-videos-COVID/6209986.html

Frenkel, S., Isaac, M., & Conger, K. (2018, October 29). *On Instagram, 11,696 examples of how hate thrives on social media*. The New York Times. Retrieved January 31, 2023, from https://www.nytimes.com/2018/10/29/technology/hate-on-social-media.html.

Gallagher, A., Hart, M., & O'Connor, C. (2021, October 20). *Ill advice: A case study in Facebook's failure to tackle COVID-19 disinformation*. ISD. Retrieved January 30, 2023, from https://www.isdglobal.org/isd-publications/ill-advice-a-case-study-in-facebooks-failure-to-tackle-COVID-19-disinformation/

Gillespie, T. (n.d.). *Content Moderation, AI, and the Question of Scale*. Sage Journals. Retrieved January 24, 2023, from https://journals.sagepub.com/doi/full/10.1177/2053951720943234

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

Google. (n.d.). *Community Guidelines Strike Basics on YouTube - YouTube Help*. Google. Retrieved January 19, 2023, from https://support.google.com/youtube/answer/2802032?hl=en

Google. (n.d.). *Harmful or dangerous content policies - YouTube Help*. Google. Retrieved January 21, 2023, from https://support.google.com/youtube/answer/2801964?hl=en&ref_topic=9282436#zippy=%2Cage-restricted-content%2Cextremely-dangerous-challenges

Google. (n.d.). Hate speech policy - *YouTube Help*. Google. Retrieved February 17, 2023, from https://support.google.com/youtube/answer/2801939?hl=en

Google. (n.d.). *Misinformation policies - YouTube Help*. Google. Retrieved February 1, 2023, from https://support.google.com/youtube/answer/10834785?hl=en

Google. (n.d.). *Violent extremist or criminal organizations policy - YouTube Help*. Google. Retrieved February 17, 2023, from https://support.google.com/youtube/answer/9229472?hl=en&ref_topic=9282436

Google. (n.d.). *YouTube Partner Program Availability - YouTube help*. Google. Retrieved February 17, 2023, from https://support.google.com/youtube/answer/7101720

Google. (n.d.). *YouTube Partner Program Overview & Eligibility - Computer - YouTube help*. Google. Retrieved February 17, 2023, from https://support.google.com/youtube/answer/72851

Gorwa, R., Binns, R., & Katzenbatch, C. (2020, February 28). *Algorithmic content moderation: Technical and political ... - sage journals*. Sage Journals. Retrieved February 20, 2023, from https://journals.sagepub.com/doi/full/10.1177/2053951719897945

Greer, R., & Ramalingam, V. (2020, February 27). The search for extremism: Deploying the redirect method. The Washington Institute. Retrieved February 26, 2023, from https://www.washingtoninstitute.org/policy-analysis/search-extremism-deploying-redirect-method

Grierson, J., Milmo, D., & Farah, H. (2021, October 8). *Revealed: Anti-vaccine Tiktok videos being viewed by children as young as nine*. The Guardian. Retrieved February 2023, from https://www.theguardian.com/technology/2021/oct/08/revealed-anti-vaccine-tiktok-videos-viewed-children-as-young-as-nine-COVID

Hagen, L. (2022, December 1). Antisemitism is on the rise, and it's not just about Ye. *NPR.org*. Retrieved from https://www.npr.org/2022/11/30/1139971241/anti-semitism-is-on-the-rise-and-not-just-among-high-profile-figures

Halprin, M., & Flannery O'Connor, J. (2022, December 1). *On policy development at YouTube*. blog.youtube. Retrieved February 1, 2023, from https://blog.youtube/inside-youtube/policy-development-at-youtube/

Han, E. (2019, August 16). *Our continued fight against hate and harassment*. Newsroom. Retrieved January 2023, from https://newsroom.tiktok.com/en-us/our-continued-fight-against-hate-and-harassment

Helft, M. (2006, October 12). *San Francisco Hedge Fund invested in YouTube*. The New York Times. Retrieved February 1, 2023, from https://www.nytimes.com/2006/10/12/technology/12hedges.html

Helft, M., & Richtel, M. (2006, October 10). *Venture firm shares a YouTube jackpot*. The New York Times. Retrieved February 1, 2023, from https://www.nytimes.com/2006/10/10/technology/10payday.html

Helmus, T., & Klein, K. (2018). Assessing outcomes of online campaigns Countering Violent Extremism: A Case Study of the redirect method. RAND Corporation Research Report Series. https://doi.org/10.7249/rr2813

Hern, A. (2021, March 9). *Instagram led users to COVID misinformation amid pandemic – report*. The Guardian. Retrieved January 31, 2023, from https://www.theguardian.com/technology/2021/mar/09/instagram-led-users-to-COVID-misinformation-amid-pandemic-report

Hsu, T. (2022, December 28). *As COVID-19 continues to spread, so does misinformation about it*. The New York Times. Retrieved February 2023, from https://www.nytimes.com/2022/12/28/technology/COVID-misinformation-online.html

Hutchinson, A. (2021, February 3). *TikTok adds new video warning labels to stop the spread of misinformation*. Social Media Today. Retrieved February 12, 2023, from https://www.socialmediatoday.com/news/tiktok-adds-new-video-warning-labels-to-stop-the-spread-of-misinformation/594481/

Instagram. (n.d.). *Combatting misinformation on Instagram*. Instagram. Retrieved February 19, 2023, from https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram

Instagram. (n.d.). *Protecting young people on instagram: Instagram blog. About Instagram.* Retrieved February 1, 2023, from https://about.instagram.com/blog/announcements/continuing-to-make-instagram-safer-for-the-youngest-members-of-our-community

Isaac, M. (2021, October 28). *Facebook renames itself, Meta*. The New York Times. Retrieved January 29, 2023, from https://www.nytimes.com/2021/10/28/technology/facebook-meta-name-change.html#:~:text=Zuckerberg%20telegraphed%20that%20his%20company,no%20longer%20tenable%2C%20he%20said.

Jiménez Durán, Rafael, The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter (November 1, 2021). Available at SSRN: https://ssrn.com/abstract=4044098 or http://dx.doi.org/10.2139/ssrn.4044098

Jones, H. (2023) *Social Media, the paradoxical freedom of speech, and our increasingly defenseless identities,* Forbes. Forbes Magazine. Available at: https://www.forbes.com/sites/hessiejones/2023/01/01/social-media-the-paradoxical-freedom-of-speech-and-our-increasingly-defenseless-identities/?sh=3c308ce68872 (Accessed: February 19, 2023).

Kang, Hyunjin, and Chen Lou. "AI Agency vs. Human Agency: Understanding Human–AI Interactions on TikTok and Their Implications for User Engagement."

Academic.oup. Oxford Academic, August 18, 2022. https://academic.oup.com/jcmc/article/27/5/zmac014/6670985.

Koebler, J., & Lamoureux, M. (2019, June 5). *YouTube miserably fails to explain why it didn't ban Steven Crowder*. VICE. Retrieved February 1, 2023, from https://www.vice.com/en/article/3k37yk/youtube-miserably-fails-to-explain-why-it-didnt-ban-steven-crowder-for-antagonizing-carlos-maza

Kraus, R. (2020, November 19). *Facebook labeled 180 million posts as 'false' since March. Election misinformation spread anyway*. Mashable. Retrieved February 12, 2023, from https://mashable.com/article/facebook-labels-180-million-posts-false

Laub, Zachary. (2019, June 7). Hate Speech on Social Media: Global Comparisons. Retrieved from Council on Foreign Relations website: https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

Leskin, P. (2020, March 16). *YouTube warns more videos than usual could be removed as content moderation is automated amid coronavirus outbreak*. Business Insider. Retrieved January 21, 2023, from https://www.businessinsider.com/youtube-coronavirus-content-moderation-video-review-machine-learning-community-guidelines-2020-3

Liebelson, D. (2014, March 28). *Map: Here are the countries that block Facebook, Twitter, and YouTube*. Mother Jones. Retrieved February 1, 2023, from https://www.motherjones.com/politics/2014/03/turkey-facebook-youtube-twitter-blocked/

Li, H. O.-Y., Bailey, A., Huynh, D., & Chan, J. (2020). YouTube as a source of information on COVID-19: A pandemic of misinformation? *BMJ Global Health*, *5*(5). https://doi.org/10.1136/bmjgh-2020-002604

Li, Y., Guan, M., Hammond, P., & Berrey, L. E. (2021, March 1). *Communicating COVID-19 information on TikTok: a content analysis of TikTok videos from official accounts featured in the COVID-19 information hub*. Academic.oup. Retrieved February 2023, from https://academic.oup.com/her/article/36/3/261/6154696

Lin, Ying. "Tiktok Us Revenue (2021–2024)." Oberlo. Oberlo. Retrieved January 2023, from https://www.oberlo.com/statistics/tiktok-us-revenue.

Little, O., & Richards, A. (2021, October 5). *TikTok's algorithm leads users from transphobic videos to far-right rabbit holes*. Media Matters for America. Retrieved February 2023, from https://www.mediamatters.org/tiktok/tiktoks-algorithm-leads-users-transphobic-videos-far-right-rabbit-holes

Map of YouTube Availability. (2022). Retrieved February 2, 2023, from https://commons.wikimedia.org/wiki/File:YouTube_Availability.png

Meta. (2021, June 1). *How Meta's third-party fact-checking program works*. Meta. Retrieved January 30, 2023, from https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works

Nelson, G. A. (2019). *Bias in Artificial Intelligence*. North Carolina Medical Journal, 80(4), 220–222. *https://doi.org/10.18043/ncm.80.4.220*

Nguyen, S. (2019, September 4). *Google and YouTube will pay record $170 million for alleged violations of children's privacy law*. Federal Trade Commission. Retrieved February 26, 2023, from https://www.ftc.gov/news-events/news/press-releases/2019/09/google-youtube-will-pay-record-170-million-alleged-violations-childrens-privacy-law

Nowill, R. (2021, July 12). *Instagram condemned for failure to moderate racist trolling of Euro 2020 players*. Hypebeast. Retrieved February 19, 2023, from https://hypebeast.com/2021/7/euro-2020-racism-social-media

O'Connor, C. (2021). *Hatescape: An In-Depth Analysis of Extremism and Hate Speech on TikTok*. ISD Global. Retrieved February 2023, from https://www.isdglobal.org/wp-content/uploads/2021/08/HateScape_v5.pdf

Oremus, W. (2022, March 5). *Analysis | the real reason Russia is blocking Facebook*. The Washington Post. Retrieved February 8, 2023, from https://www.washingtonpost.com/technology/2022/03/05/russia-facebook-block-putin-ban-roskomnadzor/

Patel, F., & Hecht-Felella, L. (2021, February 22). *Facebook's Content Moderation Rules Are a Mess*. Brennan Center for Justice. Retrieved January 18, 2023, from https://www.brennancenter.org/our-work/analysis-opinion/facebooks-content-moderation-rules-are-mess

Paul, K., & Dang, S. (2022, December 5). Exclusive: Twitter leans on automation to moderate content as harmful speech surges. *Reuters*. Retrieved from https://www.reuters.com/technology/twitter-exec-says-moving-fast-moderation-harmful-content-surges-2022-12-03/

Perrie, S. (2022, July 14). *Influencers over the age of 30 told to stay off TikTok because they're too old*. LADbible. Retrieved February 12, 2023, from https://www.ladbible.com/news/latest-influencers-over-30-told-to-stay-off-tiktok-too-old-20220714

Published by S. Dixon, & 23, M. (2022, May 23). *Instagram Users Worldwide 2025*. Statista. Retrieved January 24, 2023, from https://www.statista.com/statistics/183585/instagram-number-of-global-users/

Reducing the spread of false information on Instagram. Help Center. (n.d.). Retrieved February 19, 2023, from https://help.instagram.com/1735798276553028

Research, G. N. E. T., & Terrorism, T. A. (n.d.). *Global internet forum to counter terrorism*. GIFCT. Retrieved February 19, 2023, from https://gifct.org/

Richards, A. (2022, July 18). *Examining white supremacist and militant accelerationism trends on TikTok* . GNET. Retrieved February 2023, from https://gnet-research.org/2022/07/18/examining-white-supremacist-and-militant-accelerationism-trends-on-tiktok

Richard Delgado; Jean Stefancic, "Hate Speech in Cyberspace," Wake Forest Law Review 49, no. 2 (2014): 319-344

Roth, Y. (2022, May 22). *Introducing our crisis misinformation policy*. Twitter. Retrieved February 12, 2023, from https://blog.twitter.com/en_us/topics/company/2022/introducing-our-crisis-misinformation-policy

Sayce, D. (2010, March 3). Number of tweets per day? | David Sayce. Retrieved from David Sayce website: https://www.dsayce.com/social-media/tweets-day/

Segal, E. (2021, October 4). *Facebook responds to allegations on '60 minutes' that it contributed to Capitol Riot*. Forbes. Retrieved January 29, 2023, from https://www.forbes.com/sites/edwardsegal/2021/10/03/facebook-responds-to-allegations-on-60-minutes-that-it-contributed-to-capitol-riot/?sh=52c7097c4f42

Shang, Lanyu, Ziyi Kou, Yang Zhang, and Dong Wang. "A Multimodal Misinformation Detector for COVID-19 Short ... - IEEE Xplore." IEEE Xplore . IEEE, December 15, 2021. https://ieeexplore.ieee.org/abstract/document/9671928/.

Shead, S. (2020, November 13). *TikTok is luring Facebook moderators to fill new trust and Safety Hubs*. CNBC. Retrieved February 16, 2023, from https://www.cnbc.com/2020/11/12/tiktok-luring-facebook-content-moderators.html

Smith-Schoenwalder, C. (2020). CDC: Some People Did Take Bleach to Protect From Coronavirus. Retrieved from US News & World Report website: https://www.usnews.com/news/health-news/articles/2020-06-05/cdc-some-people-did-take-bleach-to-protect-from-coronavirus

Social Media : Misinformation and Content Moderation Issues For Congress. Congress. (n.d.). Retrieved January 24, 2023, from https://crsreports.congress.gov/product/pdf/R/R46662

Staff, T. O. I. (2022). Kanye West says he'll go to "death con 3 on JEWISH PEOPLE" after Instagram ban. Retrieved from www.timesofisrael.com website:

https://www.timesofisrael.com/kanye-west-says-hell-go-to-death-con-3-on-jewish-people-after-instagram-ban/

Stockpole, T. (2022, November 9). *Content moderation is terrible by design*. Harvard Business Review. Retrieved February 27, 2023, from https://hbr.org/2022/11/content-moderation-is-terrible-by-design

Suderman, A., & Goodman, J. (2021, October 22). *Amid the capitol riot, Facebook faced its own insurrection*. AP NEWS. Retrieved February 16, 2023, from https://apnews.com/article/donald-trump-technology-business-social-media-media-87cc5087fc653539811ee87ed79464f5

Tarasov, K. (2021, February 27). *Why content moderation costs billions and is so tricky for Facebook, Twitter, YouTube and others*. CNBC. Retrieved February 26, 2023, from https://www.cnbc.com/2021/02/27/content-moderation-on-social-media.html

TikTok . (n.d.). *User Safety*. TikTok. Retrieved February 17, 2023, from https://support.tiktok.com/en/safety-hc/account-and-user-safety/account-safety

Timberg, C., Dwoskin, E., & Albergotti, R. (2021, October 29). *Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs*. The Washington Post. Retrieved January 29, 2023, from https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/

Twitter. (n.d.). *Understanding age screening and how it works | Twitter help*. Twitter. Retrieved February 12, 2023, from https://help.twitter.com/en/safety-and-security/age-verification#:~:text=If%20you%20have%20already%20entered,on%20your%20profile%20page%20settings.

*Twitter Inc. Stock Major Holders (*2022 October 2). Yahoo Finance https://finance.yahoo.com/quote/TWTR/holders/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAFVZ1x6iVeTPGuLpJ75L

U.S. Department of Justice. (n.d.). *Facebook and Violent Extremism*. Community Oriented Policing Services. Retrieved February 17, 2023, from https://cops.usdoj.gov/RIC/Publications/cops-w0737-pub.pdf

Vickery, Jacqueline Ryan, and Jen Cardenas. "'Dear Congress, Just Play Dead': Tiktok's Absurd Guide to Surviving #LOCKDOWN." AoIR, October 16, 2021. https://ojs3-prod.lib.uic.edu/ojs/index.php/spir/article/view/12257/10425.

Vincent, J. (2019, November 25). *Instagram explains how it uses AI to choose content for your explore tab*. The Verge. Retrieved February 19, 2023, from https://www.theverge.com/2019/11/25/20977734/instagram-ai-algorithm-explore-tab-machine-learning-method

Witsil, F. (2020, October 8). *Expert: Michigan 'A hotbed for militia activity,' with growing potential for violence*. Detroit Free Press. Retrieved February 25, 2023, from https://www.freep.com/story/news/local/michigan/2020/10/08/michigan-militia-wolverine-watchmen-gretchen-whitmer/5924615002/

World Population Review. (n.d.). *Facebook Users by Country 2023*. Facebook users by country 2023. Retrieved February 16, 2023, from https://worldpopulationreview.com/country-rankings/facebook-users-by-country

YouTube. (n.d.). *Content policies & community guidelines - how YouTube works*. YouTube. Retrieved January 19, 2023, from https://www.youtube.com/howyoutubeworks/our-commitments/managing-harmful-content/

YouTube. (n.d.). *YouTube Community Guidelines & Policies - How YouTube Works*. YouTube. Retrieved January 19, 2023, from https://www.youtube.com/howyoutubeworks/policies/community-guidelines/

Countries & Policies Sources

"Abhorrent Violent Material." *Attorney-General's Department*, 21 Apr. 2021. https://www.legislation.gov.au/Details/C2019A00038

Alberti, M., & Reverdosa, M. (2023, January 14). *Brazil's Supreme Court to investigate Bolsonaro over January 8 attacks*. CNN. https://www.cnn.com/2023/01/13/americas/brazil-public-prosecutor-investigate-bolsonaro-intl-hnk/index.html

Allen, A. (2018, August 17). *Bots in Brazil: The activity of Social Media Bots in Brazilian elections*. Wilson Center. https://www.wilsoncenter.org/blog-post/bots-brazil-the-activity-social-media-bots-brazilian-elections

Agrawal, A. (2022, March 2). Germany Administrative Court holds new online hate speech regulation violates EU law. Jurist. Retrieved February 18, 2023, from https://www.jurist.org/news/2022/03/germany-administrative-court-holds-new-online-hate-speech-regulation-violates-eu-law/

Arnaudo, D. (2017). *Computational Propaganda in Brazil: Social Bots during Elections.* Samuel Woolley and Philip N. Howard, Eds. Working Paper 2017.8. Oxford, UK:

Project on Computational Propaganda. 12-15. https://demtech.oii.ox.ac.uk/research/posts/computational-propaganda-in-brazil-social-bots-during-elections/

"Australia Can Now Jail Social Media Executives over Streamed Violence." *CBS News*, CBS Interactive, Apr. 2019.https://www.cbsnews.com/news/australia-social-media-law-violent-video-streaming-illegal-facebook-new-zealand/

"Australia to Force Big Tech Companies to Hand over Disinformation Data." *Euronews*, 21 Mar. 2022.7 https://www.euronews.com/next/2022/03/21/australia-will-be-able-to-force-tech-giants-to-hand-over-disinformation-data-under-new-law

"Australia's Media Regulator to Get New Powers to Crack down on Online Misinformation." *The Guardian*, Guardian News and Media, 19 Jan. 2023.https://www.theguardian.com/media/2023/jan/20/australias-media-regulator-to-be-get-new-powers-to-crack-down-online-misinformation.

"Background Paper: Human Rights in Cyberspace." *The Australian Human Rights Commission*, Sep 2013. https://humanrights.gov.au/our-work/rights-and-freedoms/publications/background-paper-human-rights-cyberspace.

BBC. (2017, May 5). *Brazil jails eight militants over Rio Olympic plot*. BBC News. https://www.bbc.com/news/world-latin-america-39813733

BBC. (2020, December 21). Halle synagogue attack: Germany far-right gunman jailed for life. BBC News. Retrieved February 18, 2023, from https://www.bbc.com/news/world-europe-55395682

Biddle, S. (2021, October 12). *Revealed: Facebook's Secret Blacklist of "Dangerous Individuals and Organizations."* The Intercept. https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/

Biller, David. (2021, October 25). *Facebook Yanks Bolsonaro Video Claiming Vaccines Cause Aids*. AP News. https://apnews.com/article/coronavirus-pandemic-technology-health-caribbean-brazil-9abd4832762e9ed405858248547955c3.

Blanchette, J., Livingston, S., Glaser, B., & Kennedy, S. (2021, January). Protecting Democracy in an Age of Disinformation. CSIS. https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/210127_Blanchette_Age_Disinformation.pdf

Brazil: Extremism and terrorism. Counter Extremism Project. (2022, June 26). https://www.counterextremism.com/countries/brazil-extremism-and-terrorism

Brazil. Worldometer. (2023, February 1). https://www.worldometers.info/coronavirus/country/brazil/

Bukovská, B. (2019). The European Commission's Code of Conduct for Countering Illegal Hate Speech Online An analysis of freedom of expression implications.

Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 5–7. https://www.ivir.nl/publicaties/download/Bukovska.pdf

Bundesgerichtshof. (2021, July 29). *Federal Court of Justice on claims against the provider of a social network that has deleted posts and blocked accounts on charges of "hate speech"*. Federal Court of Justice. Retrieved February 24, 2023, from https://www.bundesgerichtshof.de/SharedDocs/Pressemitteilungen/DE/2021/2021149.html?nn=10690868

Bundesministerium der Justiz. (2022, December 21). Netzwerkdurchsetzungsgesetz. Bundesministerium der Justiz. Retrieved February 18, 2023, from https://www.bmj.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html

Busvine, D. (2021, July 27). *Google takes legal action over Germany's expanded hate-speech law*. Reuters. Retrieved February 19, 2023, from https://www.reuters.com/technology/google-takes-legal-action-over-germanys-expanded-hate-speech-law-2021-07-27/

Cardoso, T. (2022, November 2). *Especialistas analisam Discurso de ódio e as consequências Dessa Prática*. Instituto de Estudos Avançados da Universidade de São Paulo. http://www.iea.usp.br/noticias/especialistas-analisam-discurso-de-odio-e-as-consequencias-dessa-pratica

Chang, Y.-H. (n.d.). *Hate speech and democracy: Deciding what sort of legal doctrine is best suited to hate speech regulation in Taiwan*. Digital Repository @ Maurer Law. Retrieved February 17, 2023, from https://www.repository.law.indiana.edu/etd/79/

Cho, C., & Gallo, J. (2021). *Social Media: Misinformation and Content Moderation Issues for Congress*. https://crsreports.congress.gov/product/pdf/R/R46662

"Christchurch Shooting: Gunman Tarrant Wanted to Kill 'as Many as Possible'." *BBC News*, BBC, 24 Aug. 2020. https://www.bbc.com/news/world-asia-53861456

Clifford, B. (2021). *MODERATING EXTREMISM: THE STATE OF ONLINE TERRORIST CONTENT REMOVAL POLICY IN THE UNITED STATES*. https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/Moderating%20Extremism%20The%20State%20of%20Online%20Terrorist%20Content%20Removal%20Policy%20in%20the%20United%20States.pdf

Clyburn, James. (2011). *Our American Government | Congressman James E. Clyburn*. House.gov. https://clyburn.house.gov/fun-youth/us-government

Commission Recommendation (EU) 2018/334. *On measures to effectively tackle illegal content online*. European Commission. http://data.europa.eu/eli/reco/2018/334/oj

Constitution of the Federative Republic of Brazil of 1988. Brazilian President and National Congress. http://www.planalto.gov.br/CCIVIL_03/Constituicao/Constituicao.htm#art5xliii

Council Framework Decision 2008/913/JHA. *on combating certain forms and expressions of racism and xenophobia by means of criminal law.* The Council of the European Union. http://data.europa.eu/eli/dec_framw/2008/913/oj

Criminal Code in the version published on 13 November 1998 (Federal Law Gazette I, p. 3322), as last amended by Article 2 of the Act of 22 November 2021 (Federal Law Gazette I, p. 4906).

Cryst, E., Thiel, D., Lotufo, J. B., & Gallagher, S. (2021, November 11). *Brazil election scene setter*. Stanford Internet Observatory. https://cyber.fsi.stanford.edu/io/news/brazil-election-scene-setter

"Cyber Racism Fact Sheet (2011)." *The Australian Human Rights Commission*, 2011. https://humanrights.gov.au/our-work/publications/cyber-racism-fact-sheet-2011.

Deckler, J. (October 1, 2020). *Germany's balancing act: Fighting online hate while protecting free speech.* Politico.eu. Retrieved February 18, 2023, from https://www.politico.eu/article/germany-hate-speech-internet-netzdg-controversial-legislation/

de Perdigão Lana, P., Wagner, F. R., & Rena da Silva Santarém, P. (2022, July 11). *Proposals to regulate content moderation on social media platforms in Brazil*. Internet Society. https://www.internetsociety.org/resources/doc/2022/internet-impact-brief-proposals-to-regulate-content-moderation-on-social-media-platforms-in-brazil/

Department of Infrastructure, Transport, Regional Development, Communications and the Arts. "Online Content Regulation." *Department of Infrastructure, Transport, Regional Development, Communications and the Arts*, Department of Infrastructure, Transport, Regional Development, Communications and the Arts, 1992.https://www.infrastructure.gov.au/department/media/publications/online-content-regulation

DEPARTMENT OF JUSTICE'S REVIEW OF SECTION 230 OF THE COMMUNICATIONS DECENCY ACT OF 1996. (2020, June 3). Www.justice.gov. https://www.justice.gov/archives/ag/department-justice-s-review-section-230-communications-decency-act-1996#:~:text=The%20statute%20was%20meant%20to

Directive 2000/31/EC. *On certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce)*. European Parliament and Council. http://data.europa.eu/eli/dir/2000/31/oj

Directive (EU) 2017/541. *On combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JH*. European Parliament and Council. http://data.europa.eu/eli/dir/2017/541/oj

Does spreading epidemic rumors or false information constitute a crime? 散播疫情謠言或不實訊息是否構成犯罪？. (2020, April 10).Veterans Affairs Council, R.O.C. 國軍退除役官兵輔導委員會 https://www.vac.gov.tw/cp-2196-88511-1.html

Draper, D., & Neschke, S. (2022, October 18). *Republican Midterm Agenda: Section 230, Censorship, and Big Tech | Bipartisan Policy Center*. Bipartisanpolicy.org. https://bipartisanpolicy.org/blog/republican-midterm-agenda-section-230/

Elliott, V., & Cameron, D. (2023, February 22). *The US Supreme Court Doesn't Understand the Internet*. Wired. https://www.wired.com/story/the-supreme-court-section-230-the-internet/

Elliott, V., Petrenko, G., Hilton, S., & Deck, A. (2021, May 14). *New laws requiring social media platforms to hire local staff could endanger employees*. Rest of World. Retrieved February 24, 2023, from https://restofworld.org/2021/social-media-laws-twitter-facebook/

Escritt, T., & Schepers, S. (2019, October 9). Gunman kills two in livestreamed attack at German synagogue. Reuters. Retrieved February 18, 2023, from https://www.reuters.com/article/uk-germany-shooting-idUKKBN1WO1AL

Espejel, E. Lamm, J. Llano, I. Petrucci, T. Sharma, A. (2022, October 25). *Intermediary liability frameworks for digital platforms in Latin America still a patchwork*. White & Case LLP. https://www.whitecase.com/publications/insight/latin-america-focus-fall-2022-intermediary-liability-frameworks

European Commission (2019a). *Code of Practice on Disinformation First Annual Reports – October 2019*. PDF retrieved from https://digital-strategy.ec.europa.eu/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019

European Commission (2019b). *Progress on combating hate speech online through the EU Code of conduct 2016-2019. https://commission.europa.eu/system/files/2020-03/assessment_of_the_code_of_conduct_on_hate_speech_on_line_-_state_of_play__0.pdf*

European Commission. (2022a, June 16). *2018 code of practice on disinformation*. Shaping Europe's digital future. Retrieved January 29, 2023, from https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation

European Commission. (2022b). *The Digital Services Act: Ensuring a safe and accountable online environment*. Retrieved January 17, 2023, from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en

European Commission (2022c). *The Strengthened Code of Practice Disinformation 2022*. PDF retrieved from https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

European Commission (2022d, September 17). *European Media Freedom Act: Commission proposes rules to protect media pluralism and independence in the EU*. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_5504

European Commission(2022e). *The EU Code of conduct on countering illegal hate speech online*. PDF retrieved from https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

European Union (2012). The Charter of Fundamental Rights of the European Union. In the *Official Journal of the European Union C83* (Vol. 53, p. 380). European Union.http://data.europa.eu/eli/treaty/char_2012/oj

Facts and case summary - Snyder v. Phelps. United States Courts. (n.d.). Retrieved February 15, 2023, from https://www.uscourts.gov/educational-resources/educational-activities/facts-and-case-summary-snyder-v-phelps

FitzGerald, J. (2023, January 17). *Brazil Congress: Dozens indicted over 8 January riot*. BBC News. https://www.bbc.com/news/world-latin-america-64299892

Formulating Anti-Infiltration Law制定反滲透法. (2013, July 23). Legislative Yuan立法院. https://www.ly.gov.tw/Pages/Detail.aspx?nodeid=33324&pid=191251

"Freedom of Information, Opinion and Expression." *The Australian Human Rights Commission*, 2023. https://humanrights.gov.au/our-work/rights-and-freedoms/freedom-information-opinion-and-expression#:~:text=Section%2016%20of%20the%20Human,right%20to%20freedom%20of%20expression.

Galperin, E. (2014, August 26). *Too many secrets: A court ruling spells bad news for anonymous speech in Brazil*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2014/08/too-many-secrets-court-ruling-spells-bad-news-anonymous-speech-brazil

"Galexia Internet." *Research - Article - Jones v Toben - Racial Discrimination on the Internet (October 2002)*, Galexia.https://www.galexia.com/public/research/articles/research_articles-art22.html

General Description of the Digital Intermediary Services Act. National Communications Commission. (2022, June 29). Retrieved February 2, 2023, from https://www.ncc.gov.tw/chinese/files/22062/5532_220629_1.pdf

Germany: Flawed Social Media Law. (February 14, 2018). Human Rights Watch. Retrieved February 18, 2023, from https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

Germany: Network Enforcement Act Amended to Better Fight Online Hate Speech. (2021) Library of Congress. Retrieved February 18, 2023, from https://www.loc.gov/item/global-legal-monitor/2021-07-06/germany-network-enforcement-act-amended-to-better-fight-online-hate-speech/

Gesley, J. (2021). *Germany: Network enforcement act amended to better fight online hate speech*. The Library of Congress. Retrieved February 24, 2023, from https://www.loc.gov/item/global-legal-monitor/2021-07-06/germany-network-enforcement-act-amended-to-better-fight-online-hate-speech/#:~:text=On%20June%2028%2C%202021%2C%20the,fake%20news%20in%20social%20networks.

Goldiner, D. (2023, February 3). *AOC and Marjorie Taylor Greene resume their bitter feud, George Santos weighs in*. New York Daily News. https://www.nydailynews.com/news/politics/us-elections-government/ny-ocasio-cortez-marjorie-taylor-greene-resume-feud-20230203-aubl253ly5fttnqv4jbzux57by-story.html

Goujard, Clothilde. "German Facebook ruling boosts EU push for stricter content moderation". (July 29, 2021). Politico. Retrieved February 18, 2023, from https://www.politico.eu/article/german-court-tells-facebook-to-reinstate-removed-posts

Griffiths, James. "Australia Passes Law to Stop Spread of Violent Content Online after Christchurch Massacre." *CNN*, Cable News Network, 4 Apr. 2019.https://www.cnn.com/2019/04/04/australia/australia-violent-video-social-media-law-intl/index.html

Guo, J. (2012, February). *Publications (comments/laws/suits) 出版品(著作/法律/訴訟)*. Www.taie.com.tw. https://www.taie.com.tw/tc/p4-publications-detail.asp?article_code=03&article_classify_sn=66&sn=703

Guynn, J. (2019, February 13). *If You've Been Harassed online, You're Not alone. More than Half of Americans Say They've Experienced Hate*. USA TODAY. https://www.usatoday.com/story/news/2019/02/13/study-most-americans-have-been-targeted-hateful-speech-online/2846987002/

Hate speech and hate crime. (2023, February 8). Advocacy, Legislation & Issues. Retrieved February 15, 2023, from https://www.ala.org/advocacy/intfreedom/hate

H.R.1865-115 Congress: Allow States and Victims to Fight Online Sex Trafficking Act, H.R.
1865, 115 Cong. (2017).
https://www.congress.gov/bill/115th-congress/house-bill/1865/text

Hughes, Christopher. "Social Media Use in Australia 2022." *Statista*, 3 Jan. 2023.https://www.statista.com/statistics/680201/australia-social-media-penetration/

Important policies 重要政策. (2017, April 5). Executive Yuan 行政院. https://www.ey.gov.tw/Page/5A8A0CB5B41DA11E/a11edac6-e29a-4ebc-bfb7-a1737a90bf8e

Information about recall 罷免相關資訊. (2020, May 21). Central Election Commission 中央選舉委員會. https://www.cec.gov.tw/central/cms/321

ITS Rio. (2021). Disinformation and freedom of opinion and expression Institute for Technology and Society of Rio de Janeiro submission for the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/disinformation/2-Civil-society-organisations/Institute-for-Technology-and-Society.pdf

Johnson, A., & Castro, D. (2021, February 22). *Overview of Section 230: What It Is, Why It Was Created, and What It Has Achieved*. Itif.org. https://itif.org/publications/2021/02/22/overview-section-230-what-it-why-it-was-created-and-what-it-has-achieved/

Jurecic, Q. (2022, March 15). *What Happened the Last Time Congress Amended § 230*. Lawfare. https://www.lawfareblog.com/what-happened-last-time-congress-amended-%C2%A7-230

Kern, R. (n.d.). *White House renews call to "remove" Section 230 liability shield*. POLITICO. https://www.politico.com/news/2022/09/08/white-house-renews-call-to-remove-section-230-liability-shield-00055771

Kettemann, Matthias. " Germany: Disinformation in Pandemic Times". (September 7, 2022). John Hopkins University. Retrieved February 18, 2023, from https://www.aicgs.org/2022/09/germany-disinformation-in-pandemic-times/

Klippenstein, K., & Fang, L. (2022, October 31). *Leaked Documents Outline DHS's Plans to Police Disinformation*. The Intercept. https://theintercept.com/2022/10/31/social-media-disinformation-dhs/

Law(BR)1989/7.716. Brazilian President and National Congress. https://www.planalto.gov.br/ccivil_03/leis/L7716.htm

Law(BR)1997/9.459. Brazilian President and National Congress. https://www.planalto.gov.br/ccivil_03/leis/l9459.htm

Law(BR)2014/12.965. *Marco Civil Law of the Internet in Brazil*. Brazilian President and National Congress. https://www.cgi.br/pagina/marco-civil-law-of-the-internet-in-brazil/180

Law(BR)2016/13.260. *Anti-Terrorism in Brazil*. Brazilian President and National Congress. http://www.planalto.gov.br/CCIVIL_03/_Ato2015-2018/2016/Lei/L13260.htm

Lawson, A. (2021, March 25). *Moderating online content in the United States*. ORF. Retrieved February 16, 2023, from https://www.orfonline.org/research/moderating-online-content-in-the-united-states/

Leerssen, Paddy and Tworek, Heidi. (April 19, 2019). *An Analysis of Germany's NetzDG Law*. Transatlantic Working Group. Retrieved February 18, 2023, from https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

Library of Congress. Retrieved February 18, 2023, from https://www.loc.gov/item/global-legal-monitor/2021-07-06/germany-network-enforcement-act-amended-to-better-fight-online-hate-speech/

Lima, C. (n.d.). *Biden: Tech's liability shield "should be revoked" immediately*. POLITICO. Retrieved February 16, 2023, from https://www.politico.com/news/2020/01/17/joe-biden-tech-liability-shield-revoked-facebook-100443

Lima, C. (2023, February 22). Analysis | 5 key moments from the Supreme Court's Gonzalez v. Google arguments. *Washington Post*. https://www.washingtonpost.com/politics/2023/02/22/5-key-moments-supreme-courts-gonzalez-v-google-arguments/

Lin, L. (2022, June 15). Taiwan. Reuters Institute for the Study of Journalism. Retrieved February 17, 2023, from https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/taiwan

Lubianco, J. (2022, April 12). *Fake news bill gets stuck in Brazilian Congress and it's unlikely to be voted on before the elections; remuneration proposal for journalistic organizations is a sensitive topic*. LatAm Journalism Review by the Knight Center. https://latamjournalismreview.org/articles/fake-news-brazil-payment-journalism/

Mchangama, J. (2022). (rep.). *Thoughts for the DSA: Challenges, Ideas and the Way Forward through International Human Rights Law*. Justitia. Retrieved from https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2022/04/DSA_Commentary.pdf.

Mega, Isabel. (2022, December 16). *Ministério Da Justiça Criará Coordenadoria De Direitos Digitais*. JOTA Info. https://www.jota.info/eleicoes/ministerio-da-justica-criara-coordenadoria-de-direitos-digitais-16122022.

Mello, P. (2018, October 18). *Empresários bancam campanha contra O PT Pelo whatsapp*. Folha de S.Paulo. https://www1.folha.uol.com.br/poder/2018/10/empresarios-bancam-campanha-contra-o-pt-pelo-whatsapp.shtml

Millhiser, I. (2023, February 22). *The Supreme Court is befuddled by whether Twitter is liable for ISIS's terrorism*. Vox. https://www.vox.com/politics/2023/2/22/23610608/supreme-court-twitter-taamneh-isis-terrorism-section-230-justice

Murthy, V. (2021). *Confronting Health Misinformation*. https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf

Nagra, P. (2022, July 15). Access Alert: Taiwan Releases Digital Intermediary Services Act. Access Partnership. https://accesspartnership.com/access-alert-taiwan-releases-digital-intermediary-services-act/

Nasr, J., & Carrel, P. (2020, February 21). Germany reopens hate speech, gun law debates after shisha bar killings. Reuters. Retrieved February 18, 2023, from https://www.reuters.com/article/us-germany-shooting/germany-reopens-hate-speech-gun-law-debates-after-shisha-bar-killings-idUSKBN20F1H1

NCC collects opinions from all sectors of the "Digital Intermediary Services Act" and will carefully study and amend it to build consensus in society. NCC蒐集「數位中介服務法」草案各界意見，將審慎研議調修，凝聚社會共識. National Communications Commission. (2022, August 19). Retrieved February 1, 2023, from

https://www.ncc.gov.tw/chinese/news_detail.aspx?site_content_sn=8&is_history=0&pages=0&sn_f=47907

Online Hate and Harassment: The American Experience 2021. (2022, May 3). Www.adl.org. https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2021

"Online Hate Speech." *ESafety Commissioner*, 2023. https://www.esafety.gov.au/research/online-hate-speech.

Online Hate Speech: Findings from Australia, New Zealand, and Europe. ESafety Commissioner , 2019. https://www.esafety.gov.au/sites/default/files/2020-01/Hate%20speech-Report.pdf

Online Safety Charter - Infrastructure.gov.au, 2019. https://www.infrastructure.gov.au/sites/default/files/online-safety-charter_0.pdf.

Perrigo, B. (2021, September 10). *What Brazil's new social media rules mean for the internet*. Time. https://time.com/6096704/brazil-social-media-rules/

Protection for private blocking and screening of offensive material, 47 U.S.C. § 230 (1996). https://www.law.cornell.edu/uscode/text/47/230

Regulation(EU)2021/784. *On addressing the dissemination of terrorist content online*. European Parliament and Council.http://data.europa.eu/eli/reg/2021/784/oj

Regulation (EU) 2022/2065. *On a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. European Parliament and Council. http://data.europa.eu/eli/reg/2022/2065/oj

Reno v. ACLU. (n.d.). *Oyez*. Retrieved February 1, 2023, from https://www.oyez.org/cases/1996/96-511

Ricard, J., & Medeiros, J. (2021, February 12). *Using misinformation as a political weapon: COVID-19 and Bolsonaro in Brazil: HKS Misinformation Review*. Misinformation Review. https://misinforeview.hks.harvard.edu/article/using-misinformation-as-a-political-weapon-COVID-19-and-bolsonaro-in-brazil/

Rickards, J., Fulco, M., & Quartly, J. (2019, August 19). *The battle against disinformation.* Taiwan Business TOPICS. Retrieved February 17, 2023, from https://topics.amcham.com.tw/2019/08/battle-against-disinformation/

Robertson, A. (2022, December 19). *Supreme Court will hear Section 230 challenges in February*. The Verge. https://www.theverge.com/2022/12/19/23516769/supreme-court-section-230-gonzalez-go

Rodriguez, K, and Pinho, L. (2015, March 2). *Marco Civil Da Internet: The Devil in the Detail*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2015/02/marco-civil-devil-detail.

Rojszczak, M. (2022). Online content filtering in EU law – a coherent framework or jigsaw puzzle? *Computer Law & Security Review*, *47*. https://doi.org/10.1016/j.clsr.2022.105739

S.2448-117th Congress (2021-2022): Health Misinformation Act of 2021, S.2448, 117 Cong. (2021), https://www.congress.gov/bill/117th-congress/senate-bill/2448

S.27-117 Congress: See Something, Say Something Online Act, S. 27, 117 Cong. (2021). https://www.congress.gov/bill/117th-congress/senate-bill/27

Samios, Zoe. "Government Introduces Laws to Protect Australians from Online Misinformation." *The Sydney Morning Herald*, The Sydney Morning Herald, 19 Jan. 2023. https://www.smh.com.au/business/companies/government-introduces-laws-to-protect-australians-from-online-misinformation-20230119-p5cdqg.html

Satariano, A. and Schuetze, C. F. (January 21, 2023). *Where Online Hate Speech Can Bring the Police to Your Door*. The New York Times. Retrieved February 18, 2023, from https://www.nytimes.com/2022/09/23/technology/germany-internet-speech-arrest.html

Schultheis, E. (2021, February 2). *Germany's success against COVID is being derailed by conspiracy theories*. Slate Magazine. Retrieved February 19, 2023, from https://slate.com/news-and-politics/2021/02/germany-COVID-conspiracies-misinformation-querdenker-reichsburger-far-right.html

Section 230 of the Communications Decency Act. (2019). Electronic Frontier Foundation. https://www.eff.org/issues/cda230

Social Order Maintenance Act. Laws & Regulations Database of the Republic of China (Taiwan). (1991, June 29). Retrieved February 1, 2023, from https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=D0080067

Southerl, D. (1987, July 15). *After 38 years, Taiwan lifts martial law*. The Washington Post. Retrieved February 1, 2023, from https://www.washingtonpost.com/archive/politics/1987/07/15/after-38-years-taiwan-lifts-martial-law/6ba420e6-f061-467a-9647-63858e4956b3/

Souza, C., Steibel, F., & Lemos, R. (2017). *Notes on the Creation and Impacts of Brazil's Internet Bill of Rights*. Theory and Practice of Legislation, 5(1), 73-94.

Strobl, T. (2017, September 26). *Brazil as World LGBT Murder Capital and Rio's place in the Data*. RioOnWatch. https://rioonwatch.org/?p=37249

Taiwan: Freedom on the net 2022 country report. Freedom House. (n.d.). Retrieved February 4, 2023, from https://freedomhouse.org/zh-hant/country/taiwan/freedom-net/2022#B

Taylor, E. (2020, February 20). Germany's minorities call for action after shisha bar shootings. Reuters. Retrieved February 18, 2023, from https://www.reuters.com/article/us-germany-shooting-migrants-idCAKBN20E2S8

The Communication Security and Surveillance Act - Article Content - Laws & Regulations Database of The Republic of China (Taiwan). (n.d.). Law.moj.gov.tw. Retrieved February 18, 2023, from https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=K0060044

The White House. (n.d.). *The Constitution*. The White House. https://www.whitehouse.gov/about-the-white-house/our-government/the-constitution/#:~:text=The%20First%20Amendment%20provides%20that

Turillazzi, A., Casolari, F., Taddeo, M., & Floridi, L. (2022). The Digital Services Act: An analysis of its ethical, legal, and Social Implications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4007389

Twitter, Inc. v. Taamneh. (n.d.). *Oyez*. Retrieved February 1, 2023, from https://www.oyez.org/cases/2022/21-1496

Verification Criteria 查核準則. Taiwan Fact-checking Center 台灣事實查核中心. (2022, March 29). Retrieved February 17, 2023, from https://tfc-taiwan.org.tw/about/principle

Volokh, E. (2017). *First Amendment - Related rights*. In Encyclopædia Britannica. https://www.britannica.com/topic/First-Amendment/Related-rights

Wagner, K. (2020, March 31). *Facebook, Twitter, YouTube remove posts from Bolsonaro*. Bloomberg.com. https://www.bloomberg.com/news/articles/2020-03-31/facebook-twitter-pull-misleading-posts-from-brazil-s-bolsonaro?leadSource=uverify+wall

What Does Free Speech Mean? (n.d.). United States Courts. https://www.uscourts.gov/about-federal-courts/educational-resources/about-educational-outreach/activity-resources/what-does#:~:text=Freedom%20of%20speech%20does%20not

What is a Safe Harbor? | Winston & Strawn Legal Glossary. (n.d.). Winston & Strawn. https://www.winston.com/en/legal-glossary/safe-harbor.html

Yang, Guizi. (2022, October 19). *The importance of the "process" of crime investigation可以看看你的LINE嗎？犯罪偵查「程序」的重要性*. Plain law movement 法

律白話文運動.
https://plainlaw.me/posts/%E6%A5%8A%E8%B2%B4%E6%99%BA%EF%BD%9C%E5%8F%AF%E4%BB%A5%E7%9C%8B%E7%9C%8B%E4%BD%A0%E7%9A%84line%E5%97%8E%EF%BC%9F%E7%8A%AF%E7%BD%AA%E5%81%B5%E6%9F%A5%E3%80%8C%E7%A8%8B%E5%BA%8F%E3%80%8D%E7%9A%84

Yang, J. (2020, June 29). *When did Taiwan start to face hate speech when the United States chose to boycott Facebook? 美國選擇抵制臉書，台灣何時開始面對仇恨言論?*. 新公民議會. https://newcongress.tw/?p=20116

Zhou, G. (2022, August 31). *Do you understand the main advantages and disadvantages of the draft law on the Digital Intermediary Services Act? 數位中介服務法草案主要優缺點，你看懂了嗎?*. Taiwan Association for Human Rights台灣人權促進會. Retrieved February 17, 2023, from https://www.tahr.org.tw/news/3235