IDENTIFYING, ANALYSING, ASSESSING AND MITIGATING POTENTIAL NEGATIVE EFFECTS ON CIVIC DISCOURSE AND ELECTORAL PROCESSES: A MINIMUM MENU OF RISKS VERY LARGE ONLINE PLATFORMS SHOULD TAKE HEED OF

January 2024

CIVIL LIBERTIES UNION FOR EUROPE

European Partnership for Democracy

Authors

Sofia Calabrese
digital policy manager, EPD

Orsolya Reich, PhD
senior advocacy officer, Liberties

Research Assistant

Hannah Tilsch
advocacy and research assistant, Liberties

This paper could not have been written without the generous support of Civitates.
The sole responsibility for the content lies with the authors and it may not necessarily reflect the positions of the Network of European Foundations or the Partner Foundations.

civitates

Table of contents

Executive summary

Background and Objectives

The Digital Services Act (DSA) mandates Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to conduct thorough assessments and implement mitigation measures for systemic risks that the use of their services pose, among others, to civic discourse and electoral processes (Articles 34 and 35). In anticipation of the European Commission's guidelines due in August 2024, this paper aims to feed into the discussion on how to ensure robust protection of civic discourse and electoral processes under the DSA.

Scope and Methodology

The paper focuses on obligations related to Article 34.1(c) of the DSA. It proposes that Article 34.1(c) should be seen as setting an obligation to protect the fundamental interest of the European electorate to live in stable and well-functioning democracies. Democracy is one of the foundational values of the European Union, as articulated in Article 2 of the Treaty on the European Union.

The paper offers a structured approach for identifying risks to democratic "civic discourse" and "electoral processes". Risks to civic discourse are identified through the characteristics of civic discourse conducive to a well-functioning democracy. Risks to electoral processes feature in the public discourse in a given time frame. Therefore, a temporal approach is proposed to identify them. For each type of risk identified, corresponding mitigating measures are proposed.

All parts of the paper are derived from comprehensive research, including a review of existing literature and case studies. The novelty of this paper is that it provides a structured approach to very large online platforms and the European Commission to identify potential risks.

Civic Discourse: Characteristics and a Minimum Menu of Risks to Consider

The characteristics of a civic discourse conducive to the optimal functioning of democracies are identified below.	The risks to these characteristics are identified below.	The risk potentially applies to the following categories of VLOPs and VLOSEs.
1. The civic discourse must be inclusive, pluralistic and accessible.	• Lack of inclusivity: absence of diversity	
	• Lack of inclusivity: limited accessibility	
2. The civic discourse must recognize and respect different sociopolitical viewpoints and divisions.	• Incivility: disrespectful relations between individuals	
	• Echo chambers, selective exposure to the like-minded, isolation of perspectives	
	• Polarisation/extreme views	
	• Exacerbation of conflict situations	

The characteristics of a civic discourse conducive to the optimal functioning of democracies are identified below.	The risks to these characteristics are identified below.	The risk potentially applies to the following categories of VLOPs and VLOSEs.
3. The civic discourse must show a commitment to facts and informed dialogue, and must build citizen awareness and knowledge on pertinent issues.	• Misinformation and disinformation	
4. The civic discourse must enable citizen engagement and representative attention.	• Shadow banning of civic speech by video-sharing and social media platforms	
	• Copyright removals of content used to convey political message	
	• Organised online campaigns targeting civil society	

For each type of risk identified, corresponding mitigating measures are proposed.

Electoral Processes: Characteristics and a Minimum Menu of Risks to Consider

Relevant phases and relevance are identified below.*	The risks to electoral processes are identified below.	The risk potentially applies to the below categories of VLOPs and VLOSEs.
Pre-election phase: +++ Election day: +	• Spread of contradictory electoral promises, manipulation through micro- and nanotargeting	
Pre-election phase: +++ Election day: +	• Incorrect ad identification by upload filters: mistakenly identifying non-political ads as political and vice versa	
Pre-election phase: +++ Election day: +	• Asymmetric amplification of political content from different electoral contenders	
Pre-election phase: ++ Election day: ++ Post-electoral period: ++	• High profile politicians' posts under laxer standards for being demoted or deleted	
Pre-election phase: +++ Election day: +++ Post-electoral period: +++	• Spread of false information as regards voting processes	
Pre-election phase: +++ Election day: +++ Post-electoral period: +++	• Third-party interference	

* + indicates the risks' relevancy in a given period.

Main Recommendations for VLOPs and VLOSEs

To mitigate risks to civic discourse, it is recommended that VLOPs and VLOSEs

- implement design features that encourage inclusive discussions and give visibility to marginalised voices;
- introduce simplified language options, auto-translation, dictation features and auto-generated captions to make digital content more accessible;
- establish effective systems for users to flag and report illegal or inappropriate content;
- integrate design elements that foster inclusive and less polarising discussions;
- clearly communicate rules of engagement to improve the civility of interactions among users;
- social media platforms should be designed to encourage exposure to diverse opinions;
- increase media literacy to help users recognize and avoid echo chambers;
- present to users more content that resonates with a wide range of audiences from different groups;
- make content showing positive interactions between different political groups more visible in users' feeds;
- develop algorithms that offer a balanced information diet, exposing users to a variety of viewpoints, particularly on controversial topics.

To mitigate risks to electoral processes, it is recommended that VLOPs and VLOPEs:

- stop the processing of all observed and inferred data in the targeting of political advertising;
- restrict options available for the targeting of political advertising for provided data;
- conduct an analysis of the effectiveness of automated filters to identify political ads on social media platforms and online search engines;
- ensure a minimum level of human oversight on automated filters to be able to identify mistakes;
- monitor advertising on pages in political categories more strictly;
- ensure stricter consequences for repeated violations of requirements for political advertising;
- guarantee the consistent performance of automated filters independent of an ad's language;
- proactively provide information about how to vote to contrast false information circulated online;

- protect official accounts and websites which report information about voting;
- push corrective information about voting processes to specific users affected by disinformation.

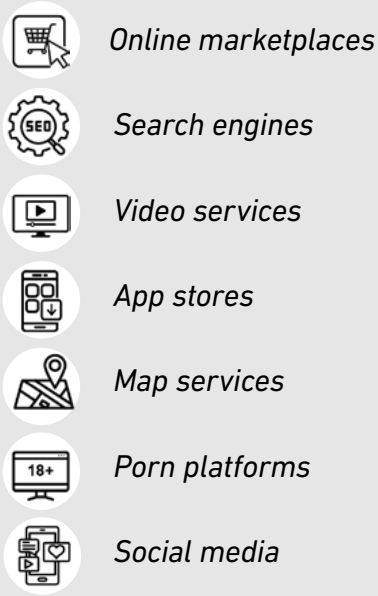
Main Recommendations for the European Commission

To ensure that civic discourse and the electoral processes are adequately protected in the European Union, the European Commission must:

- allocate adequate resources to ensure genuine implementation of Article 34.1(c);
- obtain relevant information from VLOPs and VLOSEs to better understand the dynamics behind the spread of information, the functioning of filters to identify political ads, recommender algorithms in relation to political or issue-based content;
- start an open discussion with relevant stakeholders and experts from academia and civil society;
- launch a task-force on relevant stakeholders with regular updates and analysis and the enforcement of Articles 34 and 35;
- consult with national Digital Services Coordinators about the national context of VLOP and VLOSE activities;
- publish guidelines informed by the input of relevant stakeholders such as civil society watchdogs, academia, VLOPs and VLOSEs;
- by appropriate enforcement, ensure that Article 34.1(c) and Article 35 do not become just another box-ticking exercise.

The European Commission, VLOPs and VLOSEs must also jointly assign resources to educate people to understand better how information is produced, spread, used and ordered on VLOPs and VLOSEs.

Legend



Chapter 1: Background

The Digital Services Act (DSA) is an **EU Regulation** adopted in 2022 to create a safer digital space where the fundamental rights of users are protected and to establish a level playing field for businesses. More specifically, the DSA has put forward a series of **rules applicable to different kinds of online intermediaries**, differentiated by both type and size – with obligations applicable in a cumulative way. Such obligations vary from rules and mechanisms for content moderation applicable to all sorts of online platforms, to full-fledged risk assessments to be conducted only by **Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs).**

VLOPs and VLOSEs are defined in Article 33 as "online platforms and online search engines which have a number of average monthly active recipients of the service in the Union equal to or higher than 45 million" and are designated as such by the European Commission. Currently, there are **22 designated platforms**, including social media platforms such as Facebook and X, search engines like Google Search or Bing, and online marketplaces like Amazon and Alibaba.

Among other obligations, VLOPs and VLOSEs have to **conduct risk assessments** according to Article 34 and **adopt related mitigation measures** following the criteria contained in Article 35. In particular, but not limited to, they will have to assess risks posed to "any actual or foreseeable negative effects on civic discourse and electoral processes […]".

Based on these provisions, in this paper we will first **delineate what should be understood in the context of the DSA under the terms "civic discourse" and "electoral processes", and then we will identify a series of risks and potential mitigation measures to inform VLOPs' and VLOSEs' impact assessments**, as well as the European Commission's evaluation of such assessments. We also aim to inform the European Commission's related guidelines.

Chapter 2: Definitions

Article 34.1(b) of the Digital Services Act (DSA) mandates Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to undertake risk assessments to identify, analyse and subsequently mitigate "any actual or foreseeable negative effects for the exercise of fundamental rights" enshrined in the Charter of Fundamental Rights of the European Union (EU Charter), among others the fundamental rights of "freedom of expression and information, including the freedom and pluralism of the media". Article 34.1(c) extends this obligation to "any actual or foreseeable negative effects on civic discourse and electoral processes, and public security".

EPD and Liberties are of the opinion that as the protection of fundamental rights enshrined in the EU Charter is already delineated in Article 34.1(b) of the DSA, Article 34.1(c) should be read as setting obligations that add on to what follows directly from the EU Charter. We recommend that the European Commission, as well as VLOPs and VLOSEs, interpret Article 34.1(b) as ultimately aiming to **protect the human dignity of individuals affected by the design or functioning of VLOPs' and VLOSEs' services and their related systems, and Article 34.1(c) as aiming to protect the fundamental interest of the European electorate to live in stable and well-functioning *democracies*** (cf. Recitals 81 and 82 of DSA). While living in a stable and well-functioning democracy is not an EU Charter right, democracy is one of the foundational values of the European Union, as articulated in Article 2 of the Treaty on the European Union.

As "civic discourse" and the "electoral processes" are not defined by EU law, in this chapter, we recommend interpretations of these notions based on available academic literature. In addition, we delineate the **relevant dimensions of civic discourse and the electoral processes from the perspective of digital service providers**, setting the groundwork for assessing the systemic risks posed by the services of VLOPs and VLOSEs.

I. How "civic discourse" should be understood in the context of the DSA

In recent years, European democracies have witnessed a decline. Illiberal forces advocating a purely majoritarian view of democracy have ascended to power and solidified their positions. This shift has led to a weakening of the rule of law, a foundational pillar of democratic governance. Concurrently, societal polarisation has intensified, a trend that some experts attribute to the rising influence of social media and recommender algorithms. Simultaneously, non-democratic regimes have often sought to manipulate electoral outcomes, undermining the very essence of democratic processes designed to represent the voice of the people.

Against this background, the DSA seeks to protect civic discourse, fostering a space for self-governing individuals who aim to inform themselves, form accurate beliefs, and share their understandings about socio-political developments in their respective Member States and the European Union.

Comparing the alternative delineations of the meaning of civic discourse found in academic literature and with the help of discussion of related terms, it is possible to draw insights into the ideal characteristics of democratic civic discourse. Accordingly, civic discourse is a method by which groups of people engage in reasoning about what they should do as people and who they are as people.

A conducive civic discourse for the optimal functioning of democracies:

- is inclusive, pluralistic, and accessible;

The ease with which citizens can access a variety of information, including novel viewpoints and the voices of underrepresented and marginalized communities, greatly impacts the quality of civic discourse. Civic reasoning and discourse is an inherently social endeavour, we cannot form adequate beliefs about our shared social world without the input of others. Algorithmic configurations in VLOPs and VLOSEs play a pivotal role in shaping the information landscape.

- recognises and respects differences in viewpoints and sociopolitical divisions;

In a healthy democratic setup, it's imperative to acknowledge and respect the diverse opinions, backgrounds, and identities that constitute the societal fabric. This respect for diversity forms the cornerstone of meaningful civic discourse. Online service providers must be designed in a way that discourages hateful or discriminatory rhetoric, and promotes understanding and constructive dialogue amongst members of different communities.

- shows a commitment to facts and informed dialogue, and builds citizen awareness and knowledge on pertinent issues;

A commitment to facts and informed dialogue is paramount to preserve the integrity of civic discourse. The mechanisms within VLOPs and VLOSEs to identify, flag, and counter misinformation or disinformation are crucial in maintaining an informed citizenry. The design of VLOPs, especially in terms of content amplification, moderation, and verification systems, significantly influences the quality of discourse and, by extension, the democratic process. While it is crucially important to ensure that the information citizens have access to is of good quality, VLOPs and VLOSEs should avoid over-policing content, thereby curtailing freedom of expression.

- enables citizen engagement and representative attention.

> Digital platforms serve as a bridge between citizens and their representatives, enabling a two-way communication that is critical for a responsive democratic system. The design and operational modalities of VLOPs and VLOSEs should facilitate this interaction, ensuring that citizens can effectively communicate their concerns, needs, and opinions to their representatives and vice versa. This interaction further nurtures an informed citizenry, aware of the issues at hand and capable of meaningful participation in civic discourse.

II. How "electoral processes" should be understood in the context of the DSA

According to the OSCE, "a genuine election is a **political competition** that takes place in an environment characterized by **political pluralism, confidence, transparency, and accountability.** It provides voters with an **informed choice** between distinct political alternatives. Such an election presupposes respect for basic fundamental freedoms: expression and information; association, assembly, and movement; adherence to the rule of law, including access to effective remedy; the right to freely establish political parties and compete for public office on a level playing field; non-discrimination and equal rights for all citizens, including those belonging to minority groups; freedom from intimidation and pressure; and a range of other fundamental human rights and freedoms (...)."

The **electoral process is the cornerstone of democratic governance**, enabling the citizenry to express their political will and choose their representatives. The electoral process consists of a pre-election phase, the election day itself, and a post-election phase. Across these stages, the impact of VLOPs and VLOSEs is significant. They have revolutionised how information is disseminated and consumed, which has a profound effect on politics and elections. The positive effects include increased political engagement, voter education, and the mobilisation of voters. However, the misuse of these platforms poses risks, such as the extremely fast and widespread of mis/disinformation and hate speech, which can influence voter behaviour and undermine the integrity of elections. The management of these platforms during electoral cycles is a complex task that involves balancing the benefits of open communication and information sharing against the potential for misuse that can threaten democratic processes.

In the following, we offer a **temporal approach to understanding the electoral processes:**

- pre-election phase (incl. campaign period)

> The pre-election phase has multiple dimensions: the legal framework and the election system, the election

administration, and the election campaign. VLOPs can support the pre-electoral period by providing platforms for political parties and candidates to engage with the electorate and the electorate to discuss the programs of the parties and candidates running in the elections. Voter education initiatives can also be spread effectively through social media and other digital means. **However, these platforms may also be used to disseminate disinformation or for the manipulation of public opinion through (micro- and nano-) targeted advertising campaigns**, which is especially dangerous in the days directly preceding the election day. **Incorrect political ad identification, asymmetric amplification of the content of various political contenders and third-party interference** can also put genuine political competition and the legitimacy of the results at risk.

- election day

On election day, technology, including VLOPs and VLOSEs, can be used to transmit and aggregate election results swiftly and accurately. Yet there is also the risk that real-time reporting and commentary on online platforms can lead to the spread of **false information about the voting process**, potentially causing confusion or unrest among the public.

- post-electoral period

Following the elections, VLOPs and VLOSEs can play a critical role in ensuring the transparency and credibility of the results. Conversely, they can also be a channel for the **rapid spread of claims about election fraud or other forms of post-electoral disinformation**, which can challenge the legitimacy of the process and outcomes.

Chapter 3: Risks

In this chapter, we are going through **different systemic risks to civic discourse (I) and electoral processes (II)** that we identified based on the **indications contained in Article 34.2** of the Digital Services Act regarding VLOPs and VLOSEs, in particular:

> "a) the design of their recommender systems and any other relevant algorithmic system;
> b) their content moderation systems;
> c) the applicable terms and conditions and their enforcement;
> d) systems for selecting and presenting advertisements;
> e) data related practices of the provider."

We have also identified **potential mitigating measures based on criteria contained in Article 35.1**, such as adapting the design, features or functioning of their services, including their online interfaces; adapting their terms and conditions and their enforcement; adapting content moderation processes; adapting algorithmic systems, including recommender systems; adapting advertising systems; taking awareness-raising measures; and ensuring that deepfakes are labelled as such, among others.

The following list of risks and mitigation measures is not meant to be exhaustive, but rather to present relevant examples of documented risks and mitigation measures that are susceptible to being expanded should additional evidence be available.

I. Risks to democratic civic discourse

1. Risks posed to an inclusive, pluralistic, and accessible civic discourse

Online platforms often fall short of ensuring inclusivity, leading to the marginalisation of certain voices and groups. Addressing these challenges requires a multifaceted approach. Policies that promote digital inclusion are essential. By acknowledging and actively working to bridge these gaps, online civic discourse can become more representative and accessible to a broader range of voices.

a. Lack of inclusivity: absence of diversity

As already noted in Chapter 3, the pivotal question lying at the heart of civic discourse is, "What should we do?". This question arises both in the pursuit of practical solutions and decisions, as well as in considerations of how we maintain relationships and coexist within a group.

The **lack of inclusivity and the consequent absence of diversity** in perspectives on issues of our shared social space leads to various detrimental effects, including **impaired**

decision-making, the further marginalisation of minority voices, social fragmentation, and perpetuating injustice.

Therefore, it is of utmost importance to identify its possible causes and implement measures aiming to ensure that underrepresented groups get an adequate voice so that other members of society can get to know their perspective or familiarise themselves with the specific challenges members of these groups face in our societies.

Relevant research identifies multiple risks to the availability or proper representation of certain viewpoints. **Excessive, abusive, or disproportionate content moderation, amplification, and algorithmic curation can demote or fully silence certain voices**, leading to a lack of inclusivity in online discourse.

Research indicates that feminist artists, LGBTQI+ activists, and people of colour aiming to reshape norms and challenge power structures often find their posts censored and their freedom of expression stifled online, while platforms often fail to remove hate speech and other forms of illegal expressions against the same groups. This trend often forces members of marginalised groups into self-censorship.

Demoting the voice of civil society organisations may silence those who talk about issues that have a lesser than adequate space in traditional media, such as environmental change.

To tackle these challenges, platforms can adopt **design elements that foster inclusivity.** Platforms can introduce features that nudge users towards more diverse and challenging information, thereby mitigating the risks associated with a homogenised digital public sphere. These mitigation strategies are essential for creating a more inclusive, diverse, and accessible digital public sphere where varied voices can contribute to a healthier civic discourse.

b. Lack of inclusivity: limited accessibility

Barriers to participation due to disability, language, or complexity may **hinder certain voices from participating in the public discourse and presenting their unique perspectives**. Social media platforms, integral to modern society, offer various spaces for interaction but are often not fully accessible to people with disabilities. Common issues include videos without captions, which is crucial for those who are hearing impaired, and images lacking alternative text for visually impaired users. Additionally, those with motor impairments struggle with complex platform interfaces, while individuals with cognitive disabilities face challenges in engaging with content and sharing information. Non-English speakers may struggle with participating in the global discourse, as well as people on the move or even settled immigrants who are participating in the local and national discourse. Moreover, the intricate interfaces of social media platforms pose difficulties for users with vision, motor, or cognitive impairments.

Implementing features like simple language options, dictation abilities, readability enhancements, and auto-generated captions on videos can significantly improve accessibility for users with disabilities. These mitigation

strategies are essential for creating a more inclusive, diverse, and accessible digital public sphere where varied voices can contribute to a healthier civic discourse.

Questions:

- Do current content moderation practices negatively affect the representation of marginalised voices online?
- Does algorithmic curation hinder the visibility of the viewpoints of members of different marginalised groups?
- What barriers hinder the participation of individuals with disabilities and non-English or non-native speakers in digital dialogues?

Recommendations:

- Implement design features that encourage inclusive discussions and give visibility to marginalised voices.
- Introduce simplified language options, auto-translation, dictation features, and auto-generated captions to make digital content more accessible.

2. Risks to recognizing and respecting differences and divisions in civic discourse

a. Incivility: disrespectful relations between individuals

In the digital realm, incivility presents significant challenges, stifling diverse voices and democratic engagement. Addressing these challenges requires robust reporting systems, civility-enhancing design features, and clear community guidelines to encourage positive interactions.

The **issue of incivility** in online spaces, **encompassing behaviours from impoliteness to hateful or even hate speech, significantly impacts the quality of civic discourse.** The online environment is particularly prone to such disrespectful relations between individuals, threatening democratic norms and personal freedoms. Those targeted by uncivil speech and harassment often choose to no longer discuss certain issues or opinions, or even withdraw from online discourse altogether. This negatively impacts the online space for pluralistic public opinion, as certain voices are effectively silenced, and controversial issues may not be addressed from varying perspectives. Moreover, the dissemination of hateful content has been shown to contribute to offline conflict and instability.

However, while platforms should discourage incivility among their users, **they must not conflate incivility with rude, angry, or 'overtly' emotional communicative acts.**

Anger can play a role in highlighting and emphasising the injustices underlying demands of marginalised people. Emotions can enhance the clarity with which reasoners understand that certain perspectives need to be considered earnestly, especially when these perspectives are intertwined with the personal experiences of the individuals involved in reasoning.

Civility can coexist with impolite speech or actions, particularly when such expressions are essential for conveying outrage to advance a political cause. To effectively foster civility in civic discourse, it is important to consider how one's engagement (in terms of content, form, and tone) affects others' ability to participate and to be accountable for addressing and reforming unfair interactions. Viewing civility merely as politeness can lead to the use of these norms to silence or marginalise certain voices, especially when these norms are imposed by those who did not create them or are biased towards certain participants.

Company policies on in civil speech, abuse, and harassment often lack clear definitions of punishable offences. This ambiguity has resulted in accusations that tech companies favour more powerful groups while penalising minorities. Instances of unchecked online abuse and threats against women, minorities, immigrants, refugees, and asylum seekers have been reported. Simultaneously, platforms have been accused of unjustly suppressing political dissidents, social justice activists, and those challenging oppressive structures.

Among other factors, (relative) anonymity in online spaces can exacerbate incivility through a phenomenon known as the "online toxic disinhibition effect". This effect emboldens individuals to act disrespectfully online in ways they **might not resort to in face-to-face interactions. Freedom from consequences** in online discourse can encourage incivility, including the use of insults and threatening language. At the same time, anonymity allows individuals to express opinions or present personal problems they would be afraid to express in the offline world for fear of (illegitimate) negative consequences. The challenge lies in finding ways to mitigate the toxic effect without infringing on individual rights to privacy and freedom of expression.

Such a mitigation, we believe, is possible without endangering the remains of anonymity on the Internet. Anonymity not only empowers dissidents under oppressive governments, but it also assists, for example, young people in small towns exploring their sexuality or abuse survivors embarking on a new beginning. While the harmful effects of toxic disinhibition on civic discourse must be mitigated, VLOPs may provide features enabling users to uphold civility while preserving their anonymity. On social networks, for example, individuals can have the option to manage offensive comments or block those who harass them.

Several methods can be employed to mitigate these risks. **Company policies** on uncivil speech, abuse, and harassment must be **clear, transparent, and equitably applied.** Research shows a positive association between platform design and civility. Platforms should

have **robust mechanisms** for users to report and flag inappropriate content, which is a key recommendation in protecting and promoting civic space. Affordances (what one may be able to do on a platform) can influence how users interact with each other. For example, when it takes some effort to react to other users' actions (no quote/share is available) it may foster more meaningful dialogue. Encouraging users to form discussion groups instead of platform-wide communication may also lead to more civil exchanges. Redesigning features that are often misused for negativity, for example, quote resharing, can foster a more constructive discourse. In addition, platforms could implement onboarding processes that teach and activate civility-focused norms and expectations, making them engaging and integral to the user experience.

Questions:

- What impact do online incivility and other forms of legal but potentially harmful expression have on the engagement of diverse social groups in public discourse?
- How do hate speech and other illegal content influence the quality and breadth of online discussions?
- What design and policy changes can online platforms adopt in order to enhance civility and respectful interactions?
- How can the success of these mitigation strategies be measured and evaluated over time?
- Does the platform offer users some kind of online anonymity?
- How does online anonymity contribute to the increase in uncivil behaviour in digital spaces?
- How can online platforms balance the need for civility with the protection of user privacy and freedom of expression?

Recommendations:

- Establish effective systems for users to flag and report illegal content.
- Establish effective systems for users to flag and report inappropriate content.
- Integrate design elements that foster inclusive and less polarising discussions.
- Clearly communicate rules of engagement to improve the civility of interactions among users.

b. Echo chambers, selective exposure to the like-minded, isolation of perspectives

Online echo chambers, where users engage primarily with like-minded views, are less prevalent than previously thought. Despite concerns, many users encounter diverse opinions through various media sources, and only a small segment of the population, notably those with limited media consumption who are not interested in politics, find themselves in such echo chambers. Nevertheless, this segment of the population should also be given the chance to meaningfully engage in discussions about our shared reality.

The concept of echo chambers and selective exposure on social media, primarily driven by recommender systems, has been a subject of intense scrutiny in academic research over the last decade. Echo chambers are characterised by environments where **users predominantly**

encounter opinions aligning with their own, seldom facing challenging or contrary viewpoints. Cass Sunstein, a leading legal theorist, has warned about the internet segmenting society into digital silos echoing their own voices, a process rooted in 'homophily' – the tendency to associate with similar individuals.

Recent academic insights, however, suggest that the phenomenon of echo chambers on social media may not be as widespread as was once believed. Studies indicate that only a minority of social media users find themselves in true echo chambers on a single platform. This is partly because people often access news from multiple sources, not solely from social media. Additionally, the algorithms of some platforms might not be as limiting in terms of exposure to diverse content as previously thought, with users often connecting with acquaintances outside their ideological bubbles.

Despite this, the risk of echo chambers persists, particularly among individuals who are uninterested in politics and do not consult diverse media channels.

While online echo chambers are phenomena primarily investigated in the context of social media, similar considerations may apply to other kinds of online intermediaries.

Questions:

- In what ways do social media algorithms influence the formation of echo chambers and exposure to diverse viewpoints?
- How do echo chambers affect individuals with extreme political views or limited media engagement differently from the general population?

Recommendations:

- Apply an appropriate design to encourage exposure to diverse opinions.
- Increase media literacy to help users recognize and avoid echo chambers. This involves educating users on the importance of diversifying their media sources and verifying information through multiple channels.

c. Polarisation/extreme views

VLOP and VLOSE algorithms may contribute to the rise in polarisation and the spread of extreme views. This trend poses a significant threat to democratic processes and fair electoral systems by narrowing information diversity and amplifying divisive content.

In recent years, Europe has experienced a surge in exclusionary nationalism, leading to the prominence of groups and parties that were once marginal in the political landscape. This shift is **intensified by the prevalent belief that societies** are increasingly influenced by radical opinions and **are becoming more divided.** Amidst these developments, digital technology is often identified as a contributing factor to the escalation of partisan rifts and the widening of social disparities in Europe.

Polarisation through VLOPs and VLOSEs, particularly driven by algorithmic content curation, presents a substantial risk to the

democratic process. This risk materialises as platforms, through design and moderation policies, inadvertently reduce the diversity of information accessible to individuals. This phenomenon, termed "polarisation by design", facilitates the spread of content that is not only divisive but also emotionally charged, exacerbating societal divisions and undermining free and fair discourse.

Recent studies indicate that **social media may exacerbate polarisation**, although the exact nature of this relationship remains somewhat unclear. Research shows that the degree of polarisation differs across various platforms and is influenced by the methods used to measure it. The role of online echo chambers and filter bubbles in contributing to polarisation is also not fully understood, with conflicting evidence emerging. However, a growing body of research increasingly supports the idea that social media applications are intensifying polarisation, particularly in established democracies.

Scientific studies suggest **two primary methods** to decrease polarisation. First, engaging with individuals from varied social groups, or encountering media stories about such interactions, can enhance views of these "outgroups" and lessen bias. Second, it is vital to rectify misunderstandings about the degree of polarisation, as research indicates that when discord and animosity between political factions are exaggerated, it can paradoxically intensify polarisation.

When public discourse becomes extremely polarised, research has shown that platform users tend to express fewer dissenting opinions and exhibit more withdrawal behaviours. This **self-siloing practice** not only reinforces existing echo chambers but also exacerbates polarisation, which is detrimental to an inclusive public sphere. The underlying issue stems from the algorithms' primary goal of maximising user engagement, often achieved by reducing exposure to non-alarmist content and promoting content that is shocking or even radicalising. This algorithmic approach leads to a more polarised online news knowledge base, filtered and intensified by social media platforms.

Questions:

- In what ways may VLOP and VLOSE algorithms play a role in exacerbating polarisation and the spread of extreme views?
- What are the potential long-term societal effects of continued online exposure to polarised content?

Recommendations:

- Present users with more content that resonates with a wide range of audiences from different groups.
- Make content showing positive interactions between different political groups more visible in users' feeds. This could mitigate affective polarisation, contrasting the usual virality of negative, and outrage-driven content.
- Highlight content that receives positive responses across the political spectrum.
- Alert users to content that exaggerates this divide and provides links to more accurate information.
- Do not use design features that facilitate quick negative responses to opposing

views.
- Develop algorithms that offer a balanced information diet, exposing users to a variety of viewpoints, particularly on controversial topics

d. Exacerbation of conflict situations

Exacerbation of conflict situations is a specific sub-risk that can be generated by the proliferation of hate speech online, and phenomena such as polarisation of political ideas. While the presence of online hate speech is a broader issue, the exacerbation of conflict situations is linked to specific actors and social or political tensions, and must be tackled by considering local specificities.

The exacerbation of conflict situations is a scenario **linked to the proliferation of hate speech online** (but not only) that can be caused by VLOPs when a toxic digital environment is created by hosting extremist ideologies and violent content. This poses threats to public safety and to social cohesion, creates amplified radicalisation and fuels real-world violence by providing an echo chamber for polarising and extremist views. **In conflict situations, limiting such risks is even more crucia**l, as the online debate can have real-life consequences and spark violent episodes.

For this reason, risks related to hateful speech in conflict situations should be limited by using a **"conflict sensitivity" approach**. Conflict sensitivity is defined as all interventions interacting with conflict dynamics that seek to avoid aggravating conflict. For VLOPs and VLOSEs, it is understood as the ability of online platforms to understand and mitigate the risks of their operations on conflicts and human rights in areas experiencing conflict or political instability. It encompasses efforts to minimise harm, promote peace and prevent the exacerbation of conflicts through content moderation, community guidelines and algorithmic decision-making.

Unlike the general reaction to hate speech, conflict sensitivity assessments for VLOPs and VLOSEs necessitate a **deep understanding of the local context**, including root causes, dynamics, and actors involved in conflicts.

Questions:
- How do VLOPs and VLOSEs identify drivers of online and offline manifestations of conflicts?
- How do they contribute to addressing them?
- How does platform design and content moderation advance trust, agency and cohesion in times of political unrest and conflict?
- Who are the main vulnerable groups?
- What are the unintended consequences that platform design and content moderation practices can have on conflict dynamics?

Recommendations:
- Adopt a conflict sensitivity approach to minimise harm
- Deepen the understanding of local context, dynamics and actors involved in conflict situations

3. Risks to a commitment to facts and informed dialogue, risks posed to building citizen awareness and knowledge on pertinent issues: Misinformation and disinformation

While a lot can be researched further as regards to risks to informed dialogue, building citizen awareness and issue knowledge, one of the most prominent is the presence and spread of misinformation and disinformation online. The spread of misinformation and disinformation are not at all new issues, but the characteristics of the online world make it easier to circulate false information with little to no consequence and with a much broader reach. This phenomenon is enabled or worsened by a lack of media literacy and lack of trust in institutions and can eventually lead to news avoidance. Circulation of fake news online can therefore cause polarisation in the discussion, stifling citizens' commitment to facts and informed dialogue, as well as citizen awareness and knowledge on pertinent issues. Eventually, this risks driving citizens away from political engagement. Different actions can be put in place to reduce the impact of such risk, including adopting clear and transparent policies, increasing cooperation with independent fact-checkers, and adjusting algorithmic design, among others.

Over the past years we have seen an **increase in false information online**. In every Member State of the European Union, at least half of respondents in a large sample say they come across fake news once a week or more. Similarly, in the US, 89% of adults indicate that they came across made-up news intended to mislead the public at least sometimes.

Content linked to the sharing of false information can be of different types, but it is usually understood under the umbrella of **misinformation** (false information shared without malicious intent) and **disinformation** (false information shared with malicious intent). False information is not a new challenge, but with the possibilities offered by the Internet, it is **spreading faster than before.**

The shape and spread of misinformation and disinformation are **facilitated by social media structures.** In particular, the two core attributes that create the conditions for the spread of misinformation are algorithms that promote engaging content and people's predisposition to orient towards negative news, as most "fake news" tend to evoke negative emotions. This can give rise to asymmetries in how false or misleading content and genuine content spread online, with false information arguably spreading faster and further than true information. An analysis found, for instance, that false stories circulated to a greater degree than accurate stories in the run-up to the 2016 US elections. The interpretation and classification of misleading content often involves consideration of intent and context that are difficult for third parties to assess – especially for algorithms – making it difficult to distinguish legitimate political speech from illegitimate content.

Overall, the spread of mis- and disinformation can **distort democratic engagement**, reinforce polarisation and redirect policy debates, as well as inhibit access to timely, relevant and

accurate information and data, undermining the public's willingness and ability to constructively engage in the democratic debate.

This phenomenon is enabled or worsened by a lack of media literacy and lack of trust in institutions and can eventually lead to news avoidance, as we will see in the next sections.

1. Lack of media literacy

Technological change has expanded **the range of available choices regarding access to news**. The ease of accessing news on the internet has significantly expanded opportunities for choice regarding exposure to civic discourse, but in doing so they have also **expanded the need for education on media literacy.** In a 2017 survey, only 44 percent of students 15 to 18 years old said they could identify fake news stories, and nearly one-third admitted that they had shared a story online that they later found out was inaccurate. In another survey, 84 percent of youth reported that they and their friends would benefit from instruction on how to tell if a given source of online news was trustworthy.

News and media literacy efforts are intended to **help people learn to search for, evaluate, and select online information while understanding the potential motivations, expertise, perspectives, and biases of that information**. According to the National Association for Media Literacy Education, media literacy entails "the ability to access, analyse, evaluate, create, and act using all forms of communication", and is conceived as both a way of protecting oneself against misinformation and a component of engaged, empowered civic activity. Critical media literacy goes a step further to place such reasoning within structures of power, focusing on the structures that highlight certain voices while minimising others.

The lack of media literacy **aggravates the spread of disinformation and misinformation** online and eventually leads to a lack of trust in the media, as we are going to highlight throughout the next paragraph.

2. Lack of trust in governments, media & online platforms

Additional elements that aggravate the situation regarding disinformation and misinformation online are an increasing lack of trust in governments, media, and online platforms. They can be seen as both one of the causes and consequences of the spread of fake news online.

According to a 2021 study, the significance of recent technological changes on democratic civic discourse does not stand alone but has been amplified by several broader cultural shifts, such as the **lack of trust in institutions**. Recorded trust in institutions such as national governments, the police, and news media has declined across the EU over the past two years, driven by the spread of misinformation on social media. Trust in the US government also declined from its peak in 1964 at 77 percent to less than 25 percent in the past decade. An OECD 2021 report also highlighted that across 22 countries, just over 4 in 10 people indicate trust in their national government.

The situation regarding **trust in the media** is not better. Even though public service

broadcasters are still considered the most reliable media, trust is declining in Europe, with Finland remaining the country with the highest level of trust in media (69%), while Slovakia appears to have the lowest level of trust (26%). The situation is similar in the US, where trust in mass media declined from 72 percent in 1976 to 32 percent in 2016, and half of Americans in a recent survey indicated they believe national news organisations intend to mislead or misinform. Many also expressed concern that digital platforms were at least partly to blame for declining levels of trust in news.

Finally, an increasing **lack of trust also applies to social media**. The sixth annual Insider Intelligence benchmark survey of US social media users revealed that trust in social media platforms has declined substantially in key areas, including privacy, safety, and ad relevance. A new UNICEF-Gallup study also suggests that 15- to 24-year-olds count on social media and other digital sources to stay informed, but they do not necessarily trust the information they get from them.

Distrust of institutions risks further driving away citizens who increasingly feel cynical about democratic life and participation in civic discourse because they feel that participation is inauthentic or not likely to actually influence public policy. A lack of trust in the press also leads to a **less-informed, more polarised electorate** and can be connected to behaviours such as news avoidance & news fatigue, which we will explore in the next section.

In this context, it is of utmost importance to work towards rebuilding the trust that was lost over the past years. VLOPs can play an important role in building trust in their own services, but also in news media by adopting clear & transparent policies on content moderation against disinformation and misinformation, as well as analysing data on the implementation of such policies to make sure they are correctly enforced.

3. News avoidance & news fatigue

The decreasing level of trust in media, which we just analysed, is closely correlated with a **lack of interest in news**. The increased number of news sources and accessibility of news should indicate that citizens are informed like never before. Instead, the **information overload** appears to have the opposite effect, as a 2023 Knight Foundation survey found that 61% of Americans believe these factors make it harder to stay informed.

In this context, **news avoidance** is a recently documented phenomenon where audiences reduce their consumption of journalistic media over a continuous period of time due to either an active dislike for news or a preference for other kinds of media content. A report from Reuters Institute further highlights a significant decreasing interest in news – 63% of the respondents declared being interested in news in 2017, but only 51% in 2022.

News avoidance can be **intentional** or **unintentional** and is usually caused by the perception that news coverage is too negative, by a lack of trust in the media, and by information overload. The latter can lead to a temporary state of **news fatigue** when individuals feel

overwhelmed and need to take a break. The practice of doomscrolling, that is, **obsessively checking social media feeds**, could also speed up the feeling of fatigue and is encouraged by algorithms that drive users' attention to news items that provoke emotional responses and addictive interfaces designed to maximise the time spent on the platforms.

This avoidance represents a **risk to citizens' commitment to facts and informed dialogue**, as it encourages citizens to be uninformed. Furthermore, according to a Pew Research study of nearly 15,000 citizens from 14 countries, when people withdraw from following the news about politics, they are **less likely to participate in political activities**, hence reducing political participation.

4. Spread of AI-generated deepfakes

Finally, an additional risk that has the potential to worsen the spread of disinformation and misinformation online is the rapid circulation of AI-generated deepfakes. Deepfakes are images, videos, recordings, or other types of media which are AI-generated and closely resemble real persons. For example, they could depict a politician realising false statements or an activist denying their actual positions. With deepfake technology becoming more accessible, it has been estimated that AI deepfakes have increased at an annual rate of 900% over the past few years. In this context, a survey conducted by Luminate has also found that more than half of German and French citizens are concerned about AI and deepfakes threatening election results.

The spread of disinformation and misinformation by AI-generated deepfakes can be hard to detect and can make it more difficult to find reliable information to substantiate the citizens' engagement. Furthermore, according to a report, the most advanced deepfakes can present significant threats when it comes to the spread of disinformation and misinformation, especially when they are used in key moments, such as elections.

For this reason, some platforms have put forward measures such as labelling to help distinguish between deepfakes and real content. Some have also argued that exposure and raising awareness are also possible measures to mitigate the impact of deepfakes, as well as make deepfake detection technology more accessible.

Questions:

- Are there any data available on disinformation/misinformation on the VLOPs and VLOSEs?
- How do VLOPs and VLOSEs follow their policies on disinformation?
- Does the VLOP or VLOSE use an addictive interface that could lead to doomscrolling and consequent news fatigue?
- Is space given for independent fact-checkers?
- How do recommender systems select the content to be displayed?
- Are AI-generated deepfakes present and circulated on the VLOP or VLOSE?

Recommendations:

- VLOPs and VLOSEs should adopt and

report to the Commission clear & transparent policies on content moderation against disinformation and misinformation, as well as analyse data on implementation of such policies to make sure they are correctly enforced; this would also help increase trust in online platforms and media if more quality is allowed for news items;

- VLOPs and VLOSEs should also increase cooperation with independent fact-checkers to flag information as trusted flaggers; could give some funding to independent fact-checkers;
- VLOPs and VLOSEs could put forward awareness-raising measures such as guidelines to users on how to distinguish real information from fake news; this could also be used against deepfakes;
- Restrict the addictive design techniques on VLOPs and VLOSEs to limit the phenomena of news avoidance and news fatigue;
- Analyse algorithmic design and impact of recommender systems to ensure disinformation and misinformation is not amplified;
- Put in place labels to flag deepfake content as fake; in order to do that, state-of-the art detection tools will have to be used.

4. Risks to enabling citizen engagement and representative attention

a. Shadow banning of civic speech by video-sharing and social media platforms

Shadow banning on social media and video-sharing platforms poses a significant risk to civic engagement. By quietly limiting the visibility of a user's content without their knowledge or by making certain hashtags defunct, shadow banning can silence critical voices in public debates. This practice may lead to a narrowed public discourse and reduced effectiveness in civic participation.

The term 'shadow banning' originally described a deceptive form of account suspension on web forums, where users remained unaware that their content was invisible to others. Recently, its definition has broadened to encompass subtler forms of content moderation, such as discreetly delisting and downranking content and making certain hashtags defunct.

A key feature of shadow banning is its secretive nature. Users affected by such moderation are typically unaware of the actions taken against their content. This lack of transparency raises significant concerns, particularly regarding due process and the ability to contest moderation decisions.

Shadow banning can be challenging to detect due to the personalised and dynamic nature of online content visibility. These methods often go unnoticed by users because fluctuations in

content traffic can be attributed to various factors unrelated to moderation actions.

Its most significant concern lies in the invisibility it imposes on civic voices, particularly those expressing alternative or minority opinions. This reduction in the diversity of voices in public conversations fundamentally limits the scope of democratic discourse, often resulting in a homogenised dialogue where dominant narratives are amplified, and dissenting voices are marginalised. The exclusion of these critical perspectives not only impoverishes public debates but also poses a threat to the vibrancy and inclusivity of democratic societies.

Moreover, shadow banning may disproportionately impact groups like activists, civil society organisations, minorities and those challenging the status quo. This selective silencing can skew public perception, stifle necessary societal debates, and cement existing power imbalances. In addition, social media platforms can discreetly influence user opinions via shadow banning, and may elude detection.

To address these concerns through risk mitigation measures, social media platforms must adopt more transparent and accountable content moderation practices. This includes clearly informing users about the criteria and processes involved in limiting content visibility and ensuring that these practices are fair and unbiased. Regular independent audits of platform algorithms and moderation practices can also play a crucial role in ensuring fairness and preventing unjust shadow banning. These audits should be comprehensive, assessing potential biases and disparities in content moderation.

It is to be noted that while this paper focuses on the duties of care for VLOPs and VLOSEs mandated by Article 34.1(b) of DSA, and that in itself would give little room to shadow banning (in the broad, above-described sense), other articles of the DSA effectively prohibit shadow banning. Article 14 mandates that intermediaries articulate their content moderation rules within their Terms and Conditions using precise and unequivocal language. In addition, Article 17 obligates online intermediaries to furnish a detailed Statement of Reasons each time they delete or limit access to specific content. Moreover, the DSA ensures that these decisions are subject to scrutiny, allowing for recourse through internal complaint mechanisms (as per Article 20) and external dispute resolution processes (outlined in Article 21).

Questions:

- Do VLOPs and VLOSEs ensure adequate transparency and fairness in their content moderation practices?
- What measures can be implemented to protect minority and dissenting voices from being unheard?

Recommendations:

- Effectively communicate to users the specific methods and criteria used in regulating the visibility of content.
- Involve members of vulnerable communities in developing said methods and criteria.
- Undergo regular and independent evaluations of algorithms and content moderation techniques.

b. Overzealous enforcement of copyright laws

The use of copyrighted material in political messaging poses a complex challenge in the digital age. While copyright laws are designed to protect intellectual property, their application can inadvertently impact freedom of expression, especially when such material is used in a political context. This tension often manifests in the removal of content that, while infringing on copyright, serves a significant political or public interest purpose.

Overzealous enforcement of copyright laws by digital platforms can lead to the removal of content that is crucial for political discourse. This includes instances where copyrighted materials are used under exceptions and limitations for purposes such as criticism, commentary, or parody. The removal of such content can suppress vital political discourse, limiting the public's access to diverse viewpoints and critical commentary. With the use of automated upload filters that cannot differentiate between exceptions and limitations from copyright and its infringement, platforms hinder public engagement and awareness, especially on issues where visual or audio material plays a key role in conveying the message.

Questions:

- Can automated filters differentiate between legitimate political use and genuine copyright infringement?
- Is there an adequately easy and quick way for users to challenge the automated filter's (false) copyright infringement verdict?

Recommendations:

- Clearly define and communicate policies on exceptions and limitations, ensuring that legitimate uses of copyrighted material for political purposes are not unfairly penalised.
- Establish robust and quick appeal processes for users to challenge unjustified removals/inability to upload, ensuring that decisions are fair and consider the context of use.

c. Organised online campaigns targeting civil society

The online sphere created increased opportunities for malign actors to perform organised attacks against civil society with a much wider reach than in the offline world. There is also evidence of such extremist content being amplified on some platforms by recommender systems because it is polemic and hence creates more engagement. Such a phenomenon risks the silencing of both civil society organisations and citizens willing to engage in political causes.

Across the world, digital technologies are being exploited to **silence, surveil, and manipulate**

civil society. In particular, **large-scale organised online attacks** have been documented as taking many forms, including trolling, doxing, the coordinated spread of false and defamatory information, organised attacks on websites and the use of spyware or malware to target individuals, organisations, or entire communities. In some cases, governments support non-state actors to engage in a mix of online and offline attacks against critics. Another risk is represented by bad actors creating billions of fake accounts and coordinating posts with similar content in a similar timeframe, targeting people with a certain political leaning with increasingly radical content. There is also evidence of such radical content being amplified by some platforms' recommender systems because it is polemic and hence creates more engagement.

Regrettably, there are **no industry-wide shared definitions** for organised and coordinated attacks or common standards governing how such attacks should be addressed. Furthermore, the response of social media companies to state-sponsored trolling has focused on affluent Western countries, while these issues remain largely unaddressed in the rest of the world. It is, however, relevant to note that Meta has progressively developed policies on what it called **"coordinated inauthentic behaviour,"** defining it as campaigns that include groups of fake accounts and pages seeking to mislead people about who they are.

Attacks of this kind **endanger the safety and security of civil society organisations and politically engaged citizens** and should therefore be considered as a risk to citizen engagement.

As studies around this specific problem are still at an early stage, it would be important for VLOPs to work on a **CSO-specific strategic policy document recognising the need to protect civic space from organised online campaigns** and address the challenges associated with such attacks against civil society. Furthermore, VLOPs should ensure transparency about the recommender systems they deploy, and address issues linked to amplification of extremist content, if any. Finally, VLOPs should put in place effective response mechanisms and best practices to address organised online campaigns against civil society.

Questions:

- Have there been organised campaigns against civil society on the VLOPs and VLOSEs? Is there data available on the frequency of such phenomena?
- Are there any specific policies in place to contrast organised attacks?
- Do recommender systems favour radical content used to attack civil society?

Recommendations:

- Adopt a CSO-specific strategic policy document recognising the need to protect civic space from organised online campaigns
- Ensure recommender systems transparency and address issues linked to amplification of extremist content
- Put in place effective response mechanisms and best practices to address attacks against civil society

II. Risks to electoral processes

1. Spread of contradictory electoral promises, manipulation through micro- and nanotargeting

Algorithmic systems online make it possible to target political advertising to voters with extreme accuracy, to the extent that such techniques are known as micro- and nanotargeting. These practices are particularly risky ahead of elections, as they could target swing voters and deliver tailored political messages that could even be in direct contradiction to those targeting different groups. As the use of data and personal data is the most important enabler of such practices, we suggest following the recommendation of the European Data Protection Board and restricting the use of data for targeting, including inferred and observed data.

Micro- and nanotargeting are techniques widely spread online using data, including personal data, to tailor political advertising to users. In the context of political advertising they are often used to identify so-called swing voters and target them with specific messages which can contain contradictory electoral promises. They can rely on provided (actively provided), observed (passively provided), and inferred data (data generated by algorithms based on observed and provided data; human explainable). They often also use personal data and even sensitive data to identify recipients' vulnerabilities and target them.

Such manipulation of public opinion is especially dangerous when it stems from targeted advertising campaigns **in the days directly preceding the election day** as it might exploit voters' vulnerabilities to influence the outcome of the elections, and under the limited time, the chances of getting caught in time is decreased.

For these reasons, it is recommended to analyse to what extent such techniques are used and which kind of data are exploited for targeting. Additionally, in order to mitigate the risks to the political processes, VLOPs and VLOSEs need to consider **stopping the processing of all observed and inferred data in the targeting of political advertising** – in line with the European Data Protection Board guidelines on the targeting of social media users – as well as restricting options available for the targeting of political ads for provided data, including age, language, general location and possibly some other provided identity features or declared interest categories.

Questions:

- How widespread are the techniques of micro- and nanotargeting on the platform?
- Which kind of data is used for microtargeting (provided, observed, inferred, personal, sensitive etc.)?

Recommendations:

- Stop the processing of all observed and inferred data in the targeting of political advertising.
- Restrict options available for the targeting of political advertising for provided data.

2. Incorrect ad identification by upload filters: mistakenly identifying non-political ads as political and vice versa

Political advertising has the potential to influence public opinion, civic discourse, and hence eventually electoral processes. For this reason, online service providers have started to self-regulate, and legislation has been put forward by institutions to regulate online political ads. However, such rules cannot be effective if political ads are not properly identified in the first place. Wrong identification of political ads can represent either the failure to identify them, with the consequence that they cannot be subject to specific rules; but also consist in overly restrictive moderation of ads that are not political, which can further threaten civic engagement online. Consequently, it is of utmost importance to ensure the effectiveness of the filters used to identify ads. To this end, VLOPs and VLOSEs should be encouraged to conduct an analysis of the effectiveness of their automated filters to identify political ads.

By its nature, political advertising has the potential to influence public opinion, civic discourse, and hence eventually, electoral processes. More specifically, online political advertising has a much wider reach than via traditional channels and can be delivered using potentially risky techniques, such as targeting based on personal data and microtargeting. This makes it particularly easy for malign actors to abuse ads to steer public opinion. For this reason, many online service providers have developed self-regulatory policies that include verifying advertisers' identity, creating archives of political ads or banning political ads altogether. Given the large number of submitted ads, online service providers usually deploy **automated filters for ad reviews**, complemented by human review only in some cases.

Even when legislation has been drafted to regulate online political advertising, such as in the EU's Regulation on Transparency and Targeting of Political Advertising (TTPA), an important role will still be played by online service providers. While the TTPA places emphasis on the role of the sponsor of political ads – who is obliged to declare their advertisement as political – VLOPs should still be obliged, based on the DSA's provisions, to assess the risk of sponsors not declaring that the ads are political. **The use and effectiveness of automated filters to detect political ads would, therefore, still be relevant to complement the implementation and enforcement of the TTPA rules.**

A baseline requirement for VLOPs and VLOSEs to comply with (self- and co-) regulation is, therefore, to properly identify political advertising so that internal policies or existing rules can be adequately implemented and enforced. In this context, however, there is increasing evidence of poor identification of political ads by automated filters leading to both over-, -under and mis-identification. In a 2022 study titled **An audit of Facebook's political ad policy enforcement**, a comprehensive large-scale analysis of 4.2 million political and 29.6 million non-political ads from 215,030 advertisers, has been conducted. Based on the findings of this study, it seems

that 61% more ads are missed than are detected worldwide, and 55% of U.S. detected ads are in fact non-political.

There are, therefore, two categories of incorrect identification of ads which consist of a failure of identifying political ads on one side and an over-identification of political ads on the other side. We will go through these two cases and related risks in the next paragraphs.

1. Failure to identify ads as political

Failure to identify political ads generates an opportunity for advertisers to evade restrictions on political ads, and, therefore, spread violating content to steer the online discourse and unduly influence political processes. For example, if a candidate in an election does not indicate that their ads are political and the filters miss identifying such ads as political, they would be treated as regular ads and therefore escape more stringent transparency requirements.

2. Identification of non-political ads as political

Conversely, **over-identification of political ads** could restrict civic engagement as well-meaning advertisers might be disadvantaged if their ads are unduly made unavailable due to incorrect enforcement or if they comply with policies while others do not, especially when policies are unclear or ambiguous. An example of ads that could be wrongly identified as political are ads posted by civil society organisations on specific social and political issues, with the consequence that an ad from a charity advertising for new donors could be considered political and be subject to the same transparency requirements as a candidate's ad for votes.

For these reasons, it is of utmost importance for VLOPs to ensure the effectiveness of the filters used to identify ads. While the study cited above only focused on Facebook's practices, **it would be worthwhile for all VLOPs to conduct a similar analysis** to understand the main issues around identification filters, such as the error ratio and whether over- or underidentification is prevalent. This could also help identify more specific patterns, such as in the situations in which the filters make the most mistakes, and train them accordingly.

Some **immediate recommendations** based on the studies cited above could be as follows: to ensure some level of human oversight; to be able to identify mistakes of automated filters; to expand the enforcement approach to take the advertiser into account by monitoring pages in political categories more strictly; to ensure stricter consequences for repeated violations, such as temporarily restricting advertisers from running ads; and ensure consistent performance independent of an ad's language.

Finally, it is worth noticing that these risks have been treated under the category of risks posed to electoral processes, but incorrect identification of ads also poses broader risks to civic discourse outside the electoral period by potentially stifling citizens' awareness and knowledge on important public matters. An example of that is civil society issue ads categorised as political ads by ad filters outside

electoral times that might be removed as not compliant with the requirements.

Questions:

- Does the platform allow political advertising? If yes, what are the main features of an ad to be considered political?
- What are the internal policies regarding political advertising?
- Which filters are used to detect political advertising? Is the detection mostly automated?
- What is the error ratio of the filters? Is over- or under identification prevalent (or both)?
- Are there specific patterns in the filters' mistakes?
- Are there any measures in place to tackle repeated violations?

Recommendations:

- Conduct an analysis of the effectiveness of automated filters to identify political ads
- Ensure human oversight on automated filters to be able to identify mistakes
- Monitor advertising on pages in political categories more strictly
- Ensure stricter consequences for repeated violations of requirements for political advertising
- Guarantee consistent performance of automated filters independent of an ad's language

3. Spread of false information as regards voting processes

We have already extensively analysed the risks that misinformation and disinformation pose to civic discourse, and for this we refer to section I.3. On the other hand, there are specific cases that can have a more immediate impact on electoral processes, in particular when disinformation is circulated to mislead citizens regarding elections or to discourage people from voting. Similar actions should be put in place to mitigate such risks as those recommended for disinformation, but additionally, VLOPs and VLOSEs should, among others measures, proactively provide information about how to vote; protect official accounts and websites; and push corrective information to users affected by disinformation.

While misinformation and disinformation online can undermine the public's willingness and ability to constructively engage in the democratic debate, as highlighted in section I.3, there are also specific cases in which the circulation of false information online can have **direct consequences on electoral processes and election participation.**

In recent years there has been increased attention given to the influence of disinformation on elections, with the most common forms of disinformation relating to elections including the dissemination of 'fake news' in order to discredit opponents or to influence the voting process, the falsification or manipulation of polling data, the use of fake election observation and deceptive practices to suppress vote.

In particular, a significant increase of the latter has been witnessed in the variety and volume o**f voter suppression content online with deceptive practices that can confuse voters about the mechanics of voting or undermine their confidence in the integrity of the voting system.**

1. False information to confuse voters about the mechanics of voting

Both foreign and domestic bad actors engage in voter suppression efforts by creating and disseminating false information regarding the modalities of voting. Such content can, for example, mislead the electorate on modalities of participation in the electoral process like the date and location of the election; deadlines for registering to vote; procedures for requesting, completing, and returning an absentee ballot; encouraging citizens to vote in illegitimate ways, such as by text messages. They may also disseminate posts and advertisements that falsely claim that polling locations are closed or that the entire election has been delayed.

A key prerequisite of a genuine election is that voters can access the information they require to make an informed choice. In addition to information regarding political platforms and messages, this also includes "...the who, what, when, where and how" of the electoral process and polling. Thus, disinformation about how to participate in an election may prevent intended votes from being counted, in turn **lowering voter turnout**. When these efforts target specific groups, the risk is not just an overall decline in turnout but the **exclusion of a certain group's voice**.

2. False information to undermine voters' confidence in the integrity of the voting system

In recent election cycles, online service providers have also seen the spread of disinformation designed to discourage individuals from voting, such as arousing suspicions about postal voting, the latter being controlled by the government; or intimidation, designed to discourage voters from going to their polling place or otherwise casting their ballot, such as threats of violence at the polls; and reports of law enforcement action at the polls. In 2020, at the height of the COVID-19 pandemic, disinformation was used to invoke the fear that voters would be infected at the polling location. This problem has been around for a long time, but online it can be scaled up and be effective much faster and, as in the case of general disinformation, this kind of election disinformation can be even more easily spread via algorithms and generative AI.

False information of this kind can contribute to a decline in the perceived **legitimacy of the electoral process**, as well as, once again, lower voter turnout. This is especially true on the EU level, where low turnout for elections to the European Parliament may "reaffirm its image as a 'second-order' election". Finally, it also risks **eroding trust in democracy** by spreading a consistent anti-democracy narrative that the West is in decay and that democracy is not working.

While **general mitigating measures against disinformation** that we have highlighted in section I.3 would still apply to these specific

cases, such as removing or flagging fake content, some more **tailored measures for VLOPs and VLOSEs to undertake** have also been indicated. They include **VLOPs and VLOSEs** proactively providing information about how to vote; establishing and protecting official accounts and websites; amplifying trustworthy sources; and pushing corrective information to specific users affected by disinformation.

Questions:

- Is election disinformation present on the platform? If yes, how widespread is it?
- What are the measures in place to tackle it?

Recommendations:

- Adopt measures to tackle disinformation online, as highlighted in section I.3 and as per the 2022 Code of Practice on Disinformation.
- Adopt mitigation measures more specific to voting disinformation, such as:
  - Proactively provide information about how to vote to contrast false information circulated online.
  - Protect official accounts and websites that report reliable information about voting.
  - Push corrective information about voting processes to specific users affected by disinformation.

4. Asymmetric amplification of political content from different electoral contenders

Asymmetric amplification refers to the unequal coverage or promotion of political content from different electoral contenders.

The concept of asymmetric amplification of political content, predominantly observed in traditional and broadcast media, refers to unequal representation or emphasis of political views from different electoral contenders. In the realm of broadcast media, public and private outlets are bound by various national laws during election periods. These regulations typically mandate coverage of elections in public media in a fair, balanced, and impartial manner. Private outlets are allowed to be more partial, and as long as the media landscape is fairly pluralistic, such partiality does not undermine free and fair elections.

In the case of traditional media detecting asymmetric amplification is a fairly simple matter, given the personalized and ever-changing nature of what VLOP and VLOSE users see, it is much more difficult. Research shows that VLOSE search rankings can have a huge impact on election outcomes, and given that many elections are won by a small margin, biased rankings can decide election outcomes. VLOPs can give preferential algorithmic treatment to content positively depicting their favoured contender and negative treatment of their opponents. Relatedly, both can engage in digital gerrymandering – that is, in

encouraging only the likely supporters of certain candidates to vote.

While VLOPs and VLOSEs are not public service media outlets, given their power in political opinion formation, it is essential that they are held against rather strict impartiality standards. It could sway voter preferences if they decide to give preferential treatment to certain political actors. Consequently, VLOPs and VLOSEs need to implement measures to ensure that their algorithms do not unduly favour certain political viewpoints.

Questions:

- Are there asymmetric amplifications inadvertently (or with intention) built into the algorithm?
- Are there internal policies in effect to avoid digital gerrymandering?

Recommendations:

- Implement measures ensuring that algorithms do not unduly favour certain political viewpoints.
- Educate users about the nature of content amplification and its potential impact on political discourse.

5. High-profile politicians' posts under laxer standards for being demoted or deleted

High-profile politicians often benefit from less stringent content moderation on social media, as seen with posts by figures like former US President Donald Trump. This preferential treatment, often justified by the newsworthiness of their statements, poses risks to the democratic process, including the potential for inciting real-world violence and fuelling distrust in democratic institutions.

Social media platforms have often exempted political leaders from some of their content rules under the guise of newsworthiness or public interest. This has allowed some high-profile figures to disseminate potentially harmful content with relative impunity.

The asymmetric application of content standards can have serious consequences for democratic processes. Lax controls on political leaders' speeches have been linked to the incitement of real-world political violence, as seen in the storming of the Capitol in January 2021. In addition, allowing political figures to freely disseminate misinformation can reinforce public mistrust in politicians and democratic institutions, further eroding the integrity of democratic processes.

Implementing a policy of equal content moderation for all users, regardless of their political status, can help ensure that misinformation and harmful content are adequately addressed . In addition, educating the public about the

potential for misinformation from political leaders and enhancing media literacy can help citizens critically evaluate the content they consume.

Questions:

- Does the platform strike an adequate balance between the public interest in accessing political speech and the need to prevent the spread of harmful misinformation?

Recommendations:

- Implement a policy of equal content moderation for all users, regardless of their political status, which can help ensure that misinformation and harmful content are adequately addressed.
- Educate users on the potential for misinformation disseminated by political leaders.

6. Third-party interference

Third-party interference in elections represents a significant threat to the integrity of democratic processes in Europe. This interference can come from various sources, including foreign governments, private interest groups and non-state actors. One of the most prevalent tactics used by these entities to influence elections is the orchestration of disinformation campaigns aimed at influencing voter perceptions and behaviours.

Third-party interference in elections represents a significant threat to the integrity of democratic processes worldwide. This interference can come from various sources, including foreign governments, private interest groups, and non-state actors. Their methods range from cyber-attacks on election infrastructure, to illicit financial contributions to friendly candidates and to disinformation campaigns aimed at influencing voter perceptions and behaviours.

Third-party interference may undermine **public trust** in the fairness and legitimacy of election outcomes and may lead to political instability, decreased voter turnout, and increased polarisation. Successful interference can result in policy decisions that favour the interfering party rather than the genuine will of the electorate. External interference in elections represents a direct challenge to a nation's sovereignty, potentially leading to **foreign influence over domestic policy and governance**.

Educating the electorate about the tactics used in disinformation campaigns can build resilience against manipulation. Media literacy programs can empower voters to critically assess information sources.

Questions:

- How can disinformation campaigns by third parties be better detected and countered?
- How can cooperation between VLOPs, VLOSEs and government agencies and independent election monitoring bodies be improved to safeguard elections against third-party interference?

Recommendations:

- Educate users on the potential for misinformation disseminated by third-party actors.

Chapter 4: Recommendations for the European Commission

Based on the research and evidence provided in this paper, we believe that in order to ensure that civic discourse and electoral processes are adequately protected in the European Union, the European Commission must:

- allocate adequate resources to ensure genuine implementation of Article 34.1(c);
- obtain relevant information from VLOPs and VLOSEs to better understand the dynamics behind the spread of information, the functioning of filters to identify political ads, recommender algorithms in relation to political or issue-based content;
- start an open discussion with relevant stakeholders and experts from academia and civil society;
- launch a task-force on relevant stakeholders with regular updates and analysis and the enforcement of Articles 34 and 35;
- consult with national Digital Services Coordinators about the national context of VLOP and VLOSE activities;
- publish guidelines informed by the input of relevant stakeholders such as civil society watchdogs, academia, VLOPs and VLOSEs;
- by appropriate enforcement, ensure that Article 34.1(c) and Article 35 do not become just another box-ticking exercise.

Chapter 5: Conclusions

In this paper we have worked on the practical implementation of the rules contained in Articles 34 and 35 of the EU's Digital Services Act, which mandates Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to conduct **risk assessments and implement related mitigation measures**, focusing specifically on risks to **civic discourse and electoral processes**.

Based on our definitions of civic discourse and electoral processes, we have **identified a number of risks and related mitigation measures** and split them into different categories. This allowed us to come up with a substantial but **non-exhaustive list of risks** that encompassed topics such as disinformation, political advertising, and inclusivity online. We have also put forward **recommendations for suitable mitigation measures**, such as changes in design features, in particular for algorithmic systems, content moderation practices, and enhanced data protection, along with many more recommendations specific to the different cases.

Following this methodology, we showed on one hand that there is **enough research-based evidence to identify both risks and mitigation measures**, but at the same time, we also found that **much of the existing work poses additional questions** that could only be answered by researching further into the underlying online dynamics with additional data provided by VLOPs and VLOSEs. For this reason, our list is non-exhaustive and is meant to be complemented with any additional evidence available from researchers, the European Commission, civil society, and the VLOPs and VLOSEs themselves.

To this end, we recommend the European Commission to **keep a discussion open with relevant stakeholders involved** to further clarify the functioning of certain features and tools online, their related risks, and the feasibility and effectiveness of the proposed mitigation measures. In this context, it will also be fundamental to assign adequate resources to educate people to better understand how information is produced and spread on VLOPs and VLOSEs.

Finally, it will be crucially important for the European Commission to focus on the implementation and enforcement by **publishing clear guidelines for risk assessments** based on stakeholder input and research; and to **allocate enough resources to enforce them** to make sure that efficient mitigation measures are actually put in place and the risk assessments do not become just another box-ticking exercise.

About European Partnership for Democracy (EPD)

The European Partnership for Democracy (EPD) is a network of democracy support organisations with a global remit to support democracy. Headquartered in Brussels, EPD's mission is to support democracy in Europe and around the world through the collective knowledge and capacities of European democracy support organisations.

Website

epd.eu

European Partnership for Democracy (EPD)

Rue Froissart 123-133
B-1040 Brussels
Belgium

@EPDeu, European Partnership for Democracy (EPD)

About Liberties

The Civil Liberties Union for Europe (Liberties) is a non-governmental organisation promoting the civil liberties of everyone in the European Union. We are headquartered in Berlin and have a presence in Brussels. Liberties is a network of 19 national civil liberties NGOs from across the EU.

Website

liberties.eu

The Civil Liberties Union for Europe e. V.

Ringbahnstr. 16-20
12099 Berlin
Germany

Follow us

Reference link to study

Please, when referring to this study, use the following web address:
https://www.liberties.eu/f/mpdgy5