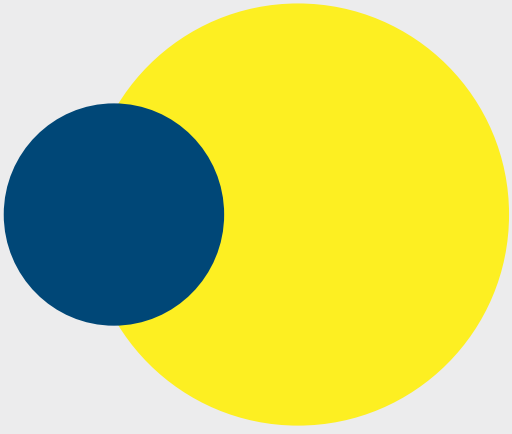
TOWARDS MEANINGFUL FUNDAMENTAL RIGHTS IMPACT ASSESSMENTS UNDER THE DSA

European Center for Not-for-Profit Law

accessnow

TABLE OF CONTENTS



INTRODUCTION

As of 25 August 2023, very large online platforms (VLOPs) and very large online search engines (VLOSEs) are required by the EU's Digital Services Act (DSA) to assess systemic risks stemming from their services. This mechanism is essential for understanding and mitigating fundamental rights risks. However, implementing the DSA's obligations for VLOPs and VLOSEs, and ensuring their effectiveness in practice, remains a challenge.

This policy paper, prepared by the European Center for Not-for-Profit Law (ECNL) and Access Now, specifically focuses on implementing Article 34(1)b of the DSA, which addresses the "actual or foreseeable negative effects for the exercise of fundamental rights." For the purpose of this paper, we use the term "fundamental rights impact assessment" (FRIA) to describe a process for implementing this Article. We argue that while Article 34 of the DSA adopts a risk-based approach, this broader framework should be grounded in fundamental rights safeguards, which are binding standards codified in the EU Charter of Fundamental Rights.

The DSA does not include provisions for a specific delegated act or guidelines that would set harmonised rules for risk assessments. Similarly, while methodologies for FRIAs have been developed by intergovernmental organisations or other stakeholders with relevant expertise,[1] there is still no consensus on what constitutes a high-quality and meaningful impact assessment. This policy paper does not propose a new parallel compliance mechanism, but rather builds on existing human rights impact assessment methodologies, calling for their harmonisation across the EU.

Our recommendations are intended to help the European Commission, and specifically its newly-established DSA enforcement team, ensure that VLOPs and VLOSEs adequately identify and evaluate the fundamental rights impacts stemming from their services. That said, our recommendations can also serve as a guide for VLOPs and VLOSEs carrying out self-assessments of their compliance with the DSA – an obligation under Article 34.

Although the focus of this paper is on the systemic risks of automated content moderation for fundamental rights, with an emphasis on the right to freedom of expression and information, our recommendations, especially the procedural ones, may also be broadly applicable to assessing other systemic risks, as listed under Article 34(1)(a),(c), and (d) of the DSA.[2]

Furthermore, although the DSA's focus is on identifying and mitigating negative effects on fundamental rights in the EU, we recommend that VLOPs and VLOSEs conduct human rights impact assessments and enact this paper's recommendations around the world, in alignment with the UN Guiding Principles on Business and Human Rights (UNGPs).

When it comes to ensuring FRIAs are meaningful, effective, and more than simply an exercise in compliance, we recommend that the European Commission, and VLOPs and VLOSEs alike, focus on the following:

1 See for instance, The Danish Human Rights Institute, Human rights impact assessment of digital activities (2018), Government of the Netherlands, Fundamental Rights and Algorithmic Impact Assessment (2021), for overview see also V. Skoric, G. Sileno, S. Ghebreab, Critical Criteria for AI Impact Assessments (2023).

2 These are: the dissemination of illegal content, negative effects on civic discourse and electoral processes, and public security, negative effects in relation to gender-based violence, the protection of public health and minor and serious negative consequences for the person's physical and mental well-being.

1. GOVERNANCE

FRIAs should be governed by the EU Charter of Fundamental Rights, which encompasses and elaborates on existing human rights standards and strengthens them with strong EU enforcement powers.

2. SCOPE

FRIAs must identify *all* negative effects on *all* fundamental rights and freedoms listed in the EU Charter that VLOPs and VLOSEs' products, services, or processes may cause, contribute to, or to which their services may be directly linked.

3. DETAIL

FRIAs should contain essential information about VLOPs/VLOSEs' systems and processes, including details of the criteria and methodologies used to determine the most pressing negative effects on fundamental rights.

4. TRANSPARENCY

FRIAs must be transparent and publicly available for external stakeholders to scrutinise.

5. HARMONISATION

Every FRIA should be flexible and "fit for purpose," while following a harmonised structure, enforced by the European Commission to ensure accurate risk identification and facilitate future oversight and accountability.

6. CONSULTATION

FRIAs must be informed and shaped by regular input from external stakeholders, including civil society organisations and impacted communities.

This paper provides more detail on how to achieve these goals, focused in the first instance on the procedural safeguards needed to lay the ground for effective, meaningful, and comprehensive FRIAs, and in the second instance, on setting benchmarks and baselines for assessing the specific risks of automated content moderation for freedom of expression – a growing concern in the AI era.

ACKNOWLEDGEMENTS

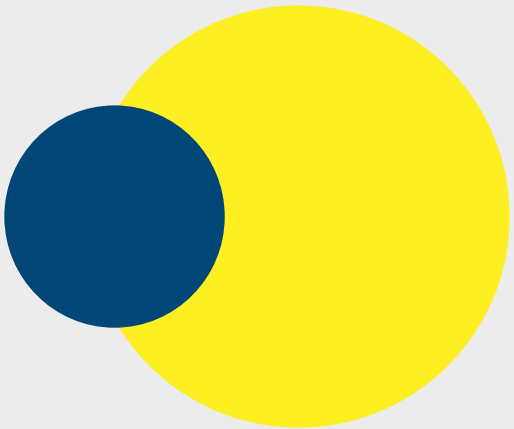
This policy paper was drafted by Eliška Pírková (Access Now), Marlena Wisniak and Karolina Iwańska (European Center for Not-for-Profit Law), with support from Estelle Massé, Fanny Hidvégi, Isedua Oribhabor, Laura Okkonen and Méabh Maguire (Access Now), and Vanja Skorič and Boglarka Szalma (European Center for Not-for-Profit Law).

Authors would like to thank the civil society organisations, companies, and individual experts who provided their input and feedback when developing the paper. In particular, we thank Business for Social Responsibility (BSR), Article One, Open Terms Archives, Anna-Katharina Meßmer from Stiftung Neue Verantwortung, Luca Belli, and Rumman Chowdhury for their time and expertise.

Prior to the publication of the paper, we made Part I.3 (stakeholder engagement) and Part II (evaluation benchmarks) available to representatives of Google, Meta, Tik Tok, Wikimedia, and Discord for their comments.

PART I. HOW TO MAKE FRIAS MEANINGFUL: SETTING SAFEGUARDS AND MOVING BEYOND A "TICK BOX" EXERCISE

The first part of this paper unpacks the main objectives of the assessment provided for in Article 34. We recommend implementing several procedural safeguards to ensure FRIAs effectively identify significant negative effects for fundamental rights, and to prevent them from becoming a performative exercise in compliance, completed by VLOPs and VLOSEs merely to avoid regulatory scrutiny. Implementing such safeguards could support FRIAs to prevent, mitigate, or remedy systemic risks to millions of people across Europe.



1. DEFINING THE SCOPE OF FRIAS

Under Article 34 of the DSA, VLOPs and VLOSEs are required to diligently "identify, analyse, and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services." To ensure that FRIAs are comprehensive and meaningful, it is essential to clarify the meaning of some of these terms.

1.1. What does "stemming from the design, functioning or use" mean?

We recommend that the notion of "stemming from the design, functioning or use" be interpreted broadly by the European Commission to encompass all negative effects that VLOPs and VLOSEs may cause, contribute to, or to which their services or use thereof may be directly linked.

To identify negative effects on fundamental rights, we recommend interpreting the sources and causes of such effects broadly, to include:

- All negative effects caused by the VLOP/VLOSE when it removes or otherwise restricts user-generated content, for example;
- All negative effects to which the VLOP/VLOSE contributed by facilitating or enabling the spread of incitement to violence and hostility or discrimination, for example; and
- All negative effects directly linked to the VLOP/VLOSE's services or use thereof, for instance, in the case of law enforcement agencies scouring the services for their own purposes.

Overall, VLOPs and VLOSEs must identify and assess all potential and actual negative effects on fundamental rights, not only those caused by their actions or omissions.

It can be challenging to establish a direct causal link between VLOPs and VLOSEs' services and the negative effect they have on societies. Platforms have denied access to public research data that would enable evidence and research-based policymaking. This prevents stakeholders from analysing the direct societal consequences of VLOPs and VLOSEs' systems and processes.[3] As a result, the negative effects on fundamental rights, alongside other systemic risks listed in Article 34 of the DSA, can be difficult to quantify; and they are primarily observed by external parties in the aggregate or over the long term. While the DSA aims to address this problematic status quo by establishing mandatory data access for researchers and civil society organisations,[4] it will still take some time for this measure to be fully implemented and operationalised.

VLOPs and VLOSEs should assess all negative effects that they contribute to through the design or functioning of their systems, as well as those directly linked to their services due to how other actors — individuals, advertisers, governments, or others — use or experience their services. Such an understanding of impact attribution aligns with Principle 17(a) of the UNGPs, which would make the assessment under the DSA consistent with the broader human rights due diligence of VLOPs and VLOSEs.

3 AlgorithmWatch, DSA must empower public interest research with public data access
4 Article 40 DSA; Mozilla submission to consultation on audits

Measures seen by platforms as performing a mitigating function may also have negative effects on fundamental rights. In the context of the DSA, it is important to clarify that negative effects stem not only from the overall design of systems and technical functionalities of the platform, but also from policies and practices intended as mitigation measures (e.g. content governance systems). Meaningful FRIAs should consider all negative effects for fundamental rights linked to a platform's services, regardless of whether they stem from its core systems and functionalities, or from measures considered by the platform to have a mitigating function.

1.2. Which elements of the service should be assessed?

We recommend that the European Commission ensures that the FRIA covers the entire service, not just the functionalities which host user-generated content or enable recipient interaction.

Recital 84 of the DSA envisions that VLOPs and VLOSEs "should focus on the systems or other elements that may contribute to the risks, including all the algorithmic systems that may be relevant, in particular their recommender systems and advertising systems."

Irrespective of their size, online platforms and search engines are generally composed of different types of systems, which include but are not limited to:

- Content moderation algorithms;
- Content recommender systems;
- Advertising systems; and
- E-commerce systems.

VLOPs and VLOSEs must examine and consider the impacts that each service may have as a whole on fundamental rights. While they can and should triage components of their services to focus on those most likely to impose negative effects on fundamental rights, they should not look exclusively at one component over another, or assess only the feature that brings the VLOP/VLOSE within the DSA's scope.

For example, if a platform providing map services falls within the scope of the DSA due to a feature enabling users to post reviews, the assessment should not only cover the risks stemming from this particular feature, but rather all parts of the service that may have negative effects on fundamental rights, including "products" in their entirety together with each "technical feature."[5]

1.3. Which groups of people should be considered in the assessment?

We recommend that the European Commission verify that the assessment identifies and analyses the negative effects on fundamental rights not only of recipients of the service — i.e. account holders or service users — but also of other potentially affected communities.

VLOPs and VLOSE's services can lead to negative effects for the fundamental rights of people who are not actively using or engaging with the service. For example, user-generated content which amounts to incitement to violence or constitutes illegal forms of hate speech can severely harm the right to life, liberty, and security of groups of people who may not be registered users of the platform where the content was distributed.

5 A product can be described as a set of characteristics enabled by the technology to serve particular functions. Products can be tangible (for example hardware), intangible (software), or a mix of both. A feature is a separate attribute, function, benefit or use that comes with a product.

In 2021, one out of five members of Europe's largest ethnic minority, the Roma minority, reported experiencing hate-motivated harassment both online and offline.[6] The widespread silencing effect of anti-gypsyism online reinforces obstacles to Roma people's participation in public life, and their use of the internet and social media.[7] Taking into account the negative effects on a wider group of affected people also aligns with existing human rights due diligence requirements and assessments under the UNGPs.

1.4. What is the main normative framework for fundamental rights compliance?

Under the DSA, the EU Charter of Fundamental Rights enables a trustworthy online environment where individuals can exercise their fundamental rights without unjustified interference by either private actors or states. The EU Charter is a relatively new and modern instrument, which applies exclusively to EU legal acts and to Member States implementing EU law, but not exclusively Member States' national laws.

The European Commission must guarantee that the DSA enforcement at both European and national levels, including all delegated acts, implementing regulations, and guidelines or risk mitigation measures, as well as Member States' national implementing acts, are fully compatible with the EU Charter of Fundamental Rights' legally binding rights and principles.

While the DSA must comply with all legally binding rights and principles enshrined in the EU Charter, EU legislators have highlighted the right to freedom of expression and information, the freedom to conduct a business, the right to non-discrimination, and the attainment of a high level of consumer protection.[8] The DSA regulates essential areas for these rights such as intermediary liability, dissemination of illegal content online, or the sale of counterfeit goods on the internet. However, the DSA and its enforcement must comply with the EU Charter as a whole.

VLOPs and VLOSEs should not therefore be able to pick and choose which rights will be considered in their assessments. While prioritising some rights over others might be necessary due to the existing negative effects of concrete systems, all negative effects on fundamental rights must first be properly mapped and explained in preliminary assessments. The prioritisation process should follow procedural safeguards, as outlined in the second part of this paper.

1.5. What is the relationship between the EU Charter and international human rights law?

All EU Charter safeguards are, as a minimum, equivalent to those included in the European Convention on Human Rights (ECHR). The direct reference to the ECHR in the Charter covers both the Convention and its protocols. In this vein, the meaning and scope of rights are also determined by European Court of Human Rights (ECtHR) case law, as well as the Court of Justice of the European Union (CJEU).[9] Together with EU Member States' national constitutions and case law, they make up a body of European human rights doctrine.

European human rights doctrine takes a similar approach to the UN International Covenant on Civil and Political Rights (ICCPR) and Economic, Social, and Cultural Rights (ICESCR), as well as the core international conventions.[10] Many of the EU Charter's articles reflect the provisions of international human rights instruments, making these highly relevant when interpreting the Charter. At an EU level, the Charter strengthens enforcement of existing human rights obligations since it benefits from the legal power of the EU law. To provide a practical example, under the ECHR, interference with freedom of expression must follow a three-step test similar to the ICCPR. Furthermore, the EU Charter sets an additional requirement for interference: to respect the essence of the right to freedom of expression.[11]

6 EU Fundamental Rights Agency, FRA's 2021 survey on Roma.
7 Minority Rights Group Europe, Freedom from hate: Empowering civil society to counter cyberhate against Roma
8 See Recital 3 of the Digital Services Act
9 EU FRA Handbook on the Charter
10 See United Nations Human Rights Office of the High Commissioner of Human Rights. The Core International Human Rights Instruments and their monitoring bodies. <https://www.ohchr.org/en/core-international-human-rights-instruments-and-their-monitoring-bodies>
11 Article 52 of the EU Charter.

VLOPs and VLOSEs need not, therefore, reinvent the wheel when creating the FRIA methodologies prescribed by the DSA. The EU Charter encompasses and elaborates on existing human rights standards, strengthening them with strong EU enforcement powers. Alongside the UNGPs, existing human rights impact assessment methodologies should serve as a baseline for FRIA compliance, and ensure they fulfil their ambition of preventing any potential, ongoing, or future significant negative effect on fundamental rights.

1.6. How does the EU Charter apply to private actors?

Once the EU Charter applies, all individuals that fall into effective jurisdiction of the EU and its Member States can rely on its protection. VLOPs and VLOSEs also benefit and are obliged to comply with the EU Charter.

Under international human rights law, the UNGPs establish companies' responsibility to respect human rights under Pillar I of its framework. Pillar I of the UNGPs reiterates States' duty to protect human rights, including to ensure that private entities under their jurisdiction fulfil this obligation. The DSA contains a direct reference to the UNGPs, calling on intermediary services to pay due regard to international human right safeguards.[12] VLOPs and VLOSEs have extensive experience with conducting Human Rights Impact Assessments, following well-established methodologies in this area. Therefore, the commonalities between the EU Charter and international human rights bills described above ensure that FIRAs under the DSA are not a new compliance mechanism but rather builds on existing best practices supported by the strong enforcement power of the EU law.

Furthermore, all EU Member States are bound by positive obligation to provide safeguards against abuse or unjustified restrictions of fundamental rights by private actors. In other words, they have to create the societal conditions for the free exercise of rights to prosper. Rules may be needed to prevent arbitrary decisions by VLOPs and VLOSEs, for instance to remove legitimate expression of speech.[13] FRIAs are a helpful tool to ensure that the legally binding force of the EU Charter governs VLOPs' and VLOSEs' systems and processes.

12 DSA, Recital 47
13 Aleksandra Kuczerawy, Safeguards for Freedom of Expression in the Era of Online Gatekeeping.

2. SETTING MINIMUM REQUIREMENTS FOR IDENTIFYING AND ASSESSING FUNDAMENTAL RIGHTS IMPACTS

In this section, we outline the minimum requirements that every FRIA process should follow, so as to evaluate the accuracy, completeness, and meaningfulness of the FRIA regardless of the methodology used. The implementation of proposed requirements by VLOPs and VLOSEs should be monitored and verified by the European Commission.

2.1. What essential components must be included in every FRIA?

1. A detailed description of the service's structure

To understand the VLOP's or VLOSE's overall architecture and the object of the FRIA, the European Commission, auditors, and other stakeholders need a detailed description of the service's structure. VLOPs and VLOSEs should describe the service and technical products that exist on the platform or search engine and map all relevant systems, including algorithmic systems, specifying their purpose and the technical functionalities they are composed of. They should also **explain how these different components build on or feed into each other and integrate into the overall service**, similarly to what the EU's draft AI Act proposes for high-risk AI systems.[14] While some VLOPs are using machine learning techniques to identify new instances of hate speech or non-consensual nudity, most are merely using a sophisticated version of pattern matching: comparing new content to a blacklist of already known examples.[15]

When describing an automated content moderation system, they should first indicate for which products or parts of the service this system is used. Second, they should indicate what "hard" content moderation systems are being used (i.e. systems that classify user-generated content based on either matching or prediction); and third, how this system interacts with "soft" moderation systems (e.g. content recommender systems).[16] As this description is crucial for understanding the scope of the assessment, VLOPs and VLOSEs should not be able to invoke the need to protect trade secrets to prevent sharing this information with auditors or the Commission under the guise of security reasons.[17]

When assessing algorithmic systems' fundamental rights impacts, we recommend that VLOPs and VLOSEs follow a similar approach to what is proposed in the EU AI Act. The European Commission should specifically ensure that VLOPs and VLOSEs provide a detailed description of the systems' key design specifications, such as the algorithmic systems' general logic, indicators of accuracy and any error rate, the rationale and assumptions made about persons or groups of persons who can be affected, the main classification choices, what the system is designed and optimised for, parameter relevance, and decisions about any possible trade-offs.

2. A detailed description of the FRIA methodology

As a minimum, VLOPs and VLOSEs should report on the indicators and scales they use in the FRIA, how they balance competing interests, and how they assess proportionality (trade-offs), answering the following questions:[18]

14 Note that in the spirit of ensuring the integrity of the EU regulatory landscape we borrow some of the terminology from the proposal for the Artificial Intelligence Act (Annex IV point 2(b)).
15 Tarleton Gillespie, Content moderation, AI, and the question of scale (2020).
16 Robert Gorwa, Reuben Binns, Christian Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance (2021).
17 European Digital Rights (EDRi) Digital Services Act: The EDRi guide to 2,297 amendment proposals (2020).
18 Based on criteria proposed by V. Skoric, G. Sileno, S. Ghebreab, Critical Criteria for AI Impact Assessments (2023).

- What specific indicators do they use to assess impacts on fundamental rights? Are these aligned with the Charter's requirements?
- What metrics and scales do they use to assess each negative effect's likelihood and severity?
- How do they assess the level of impact to each affected fundamental right?
- Do they assess how both the intended and unintended use of their services impacts fundamental rights?
- How is the proportionality and necessity of actual and potential negative effects on each impacted right balanced against a specific measure's goals?
- Which criteria are considered when deciding whether it is appropriate to continue with the development or use of a certain practice or technical functionality, given actual or potential negative effects?

3. A preliminary assessment of all negative effects on fundamental rights and freedoms, as defined in the EU Charter

We recommend listing and categorising all identified negative effects in the preliminary stage of the FRIA following the structure of each chapter of fundamental rights and principles used in the EU Charter:

- **Chapter 1: Dignity;**
- **Chapter 2: Freedoms;**
- **Chapter 3: Equality;**
- **Chapter 4: Solidarity;**
- **Chapter 5: Citizens Rights;**
- **Chapter 6: Justice.**[19]

The preliminary assessment should outline actual (i.e. existing) and foreseeable (i.e. potential or anticipated) negative effects for each cluster of fundamental rights. In identifying these effects, VLOPs and VLOSEs should consider both their own internal analyses and any publicly available reports, articles, or resources documenting existing impacts. They should also carry out regular consultations with relevant stakeholders in order to accurately identify both actual and foreseeable negative effects. This includes engaging with diverse and interdisciplinary teams and departments within the platform to identify possible risks or threats (e.g. in a red-teaming exercise[20]), and with external stakeholders, such as civil society, academia, and potentially affected community representatives, among others.

4. A list of the most significant negative effects on fundamental rights previously identified in the preliminary assessment

VLOPs and VLOSEs should determine which negative effects are the most severe and likely to occur, and explain why. A list of the most significant negative effects will help determine the systemic fundamental rights risks stemming from VLOPs and VLOSEs' systems and processes. We elaborate on criteria for the prioritisation below.

19 To see what rights and principles are included in each cluster, please consult the EU charter or the EU FRA Guidance on Applying the Charter of Fundamental Rights of the European Union in law and policymaking at national level.

20 Red-teaming is a mostly interdisciplinary process commonly adopted in the technology industry to review or challenge companies' policies, systems, and assumptions, by taking an adversarial approach (in other words, 'what can go wrong and why?').

5. An in-depth assessment of significant negative effects on fundamental rights

At a minimum, the European Commission should require VLOPs and VLOSEs to describe:

- **The purpose of each system and service that is being assessed;**
- **How they assess the significance of negative effects based on their severity, probability, number of affected persons (broken down by users and non-users), irreversibility, and possibility to remedy;**
- **How they assess necessity, i.e. whether or not the objectives can be achieved by means that do not cause negative effects on fundamental rights;**
- **How they assess proportionality, i.e. balancing whether the negative effects are the least intrusive and justified in light of the purpose of the system or service;**
- **Whether they considered any regional or linguistic aspects in the FRIA, pursuant to Article 34(2) of the DSA;**
- **Which groups of people could be negatively affected, and whether it is known (to the extent possible without collecting additional data) if they belong to marginalised communities.**

In Part II of this paper, we will develop a few hypotheses and questions to facilitate how the Commission evaluates this in-depth assessment, specifically in the context of automated content moderation.

6. Future mitigation measures or actions following the FRIA

For any negative effects that can be mitigated, platforms should describe what mitigation measures can be adopted, and explain how it addresses the fundamental rights impacts.

We recommend that VLOPs and VLOSEs explicitly indicate which negative effects they have determined to be fundamentally incompatible with the Charter. This will be the case when abuse of right occurs, e.g. when an act aims at the destruction of any rights and freedoms recognised in the Charter (the so-called "abuse of rights" clause)[21] and where there is no adequate mitigation measure. They should then discontinue the practice or technical functionality linked to the negative effect, or refrain from introducing a new practice or technical functionality.

21 Article 54 of the Charter.

2.2. How should significant negative effects on fundamental rights be prioritised?

While the DSA contains no guidelines or directives for prioritising the in-depth assessment of negative effects on fundamental rights, it may be necessary in practice to prioritise the most pressing negative effects to make the best use of limited resources.This does not mean, however, that other negative effects can be neglected altogether. Ultimately, VLOPs and VLOSEs must evaluate all systemic risks to all fundamental rights, as prescribed by the EU Charter. Our recommendations aim to merely guide their decision-making process on prioritisation. These recommendations can also guide the European Commission when reviewing VLOPs and VLOSEs' assessments of negative effects on fundamental rights. **We recommend that the following, non-exhaustive criteria be taken into account when prioritising:**

1. Significant negative effects on fundamental rights[22]

VLOPs and VLOSEs should prioritise assessing the most significant negative effects on fundamental rights. When determining the significance of potential negative effects and impacts, Recital 79[23] requires that platforms consider:

- **The severity of impact**, i.e. how seriously a specific right is impacted, based on how many people (users and non-users) are affected, whether the impact is irreversible, and whether it is possible to remedy and restore the situation to the same or equivalent position from before the impact arised, among others;
- **The probability of impact**, i.e. the likelihood that the impact will occur, based on negative effects which have previously arisen on the platform or for other companies in the sector.

2. External factors, such as a geopolitical or emergency situation

VLOPs and VLOSEs should be aware of geopolitical tensions or emergencies, and their related negative effects on fundamental rights. This means, for example, prioritising an assessment of risks related to the use of their services in conflict zones or in countries or regions with upcoming elections.

3. The risk of heightened significant negative effects on vulnerable service recipients

VLOPs and VLOSEs should take into account whether any of their content policies or content moderation systems target, or disproportionately harm, people based on characteristics protected under EU non-discrimination law, i.e. the open-ended list of grounds protected against discrimination prescribed by the EU Charter.[24] For example, TikTok demoted the content of people with disabilities, allegedly to protect them from bullying.[25] In assessing this, VLOPs and VLOSEs should take into consideration the relevant context of individual EU Member States.

4. Prior restraints on fundamental rights via automated means

Automation can exacerbate existing negative effects on fundamental rights, by restricting rights before a person can exercise them (e.g. when content is removed prior to publication), increasing the scale of negative effects, or adversely impacting access to remedy or human oversight. Therefore, VLOPs and VLOSEs should generally consider automated systems, including algorithmic systems, as being potentially linked to heightened negative

22 We understand this notion as equal to "salient adverse impacts" as understood by the Principle 24 of the UNGPs. According to Guiding Principle 24 of the UNGPs, companies should "first seek to prevent and mitigate those that are most severe or where delayed response would make them irremediable." Similar language can be found in recital 79 of the DSA.

23 DSA, Recital 79: In determining the significance of potential negative effects and impacts, providers should consider the severity of the potential impact and the probability of all such systemic risks. For example, they could assess whether the potential negative impact can affect a large number of persons, its potential irreversibility, or how difficult it is to remedy and restore the situation prevailing prior to the potential impact.

24 For further details, please consult Article 21 of the EU Charter

25 https://www.bbc.com/news/technology-50645345

effects for fundamental rights, which would justify prioritisation. For example:

- Upload filters, which automatically remove content prior to publication, typically have more significant negative impacts for fundamental rights than when content is merely flagged for human review;
- The use of natural language processing (NLP) systems or generative AI can lead to discriminatory bias, due to how these systems have been developed;
- "Shadow banning" and content demotion reduces the visibility of content in a way that can be discriminatory; and
- Profiling systems which predict, for example, the likelihood, based on past behaviour, of specific users spreading potentially harmful but legal content (e.g. mis- or disinformation) can place undue restrictions on users' freedom of expression prior to such content even being published.

2.3. How should the FRIA be structured?

In time, we recommend that the European Commission create a flexible yet harmonised structure for FRIAs. The assessment should clearly distinguish impacts related to specific features and systems of the service, to accurately identify risk and facilitate future oversight and accountability.

The assessment should be structured in a way that contextualises negative effects and links specific negative effects on fundamental rights to specific systems and their elements.

Ultimately, the person reviewing or evaluating the assessment should be able to easily understand which systems, technical functionalities, policies, or practises the negative effect stems from, and whether the negative effect is linked to one or more of these elements' specific characteristic or design features. For example, it wouldn't be enough for the VLOP or VLOSE to claim that their content moderation systems and policies may, in general, lead to an over-removal of legal content, thus impacting freedom of expression. They must clearly indicate which elements of these systems or policies – such as the algorithmic system's design features, for instance – could cause the identified negative effect and explain why.

As well as making it easier for VLOPs and VLOSEs to accurately identify risks and adopt appropriate and tailored mitigation measures, this approach also facilitates auditing, external oversight, and enforcement.

2.4. When should a new impact assessment be carried out?

As per Article 34(1) of the DSA, FRIAs shall be carried out at least once a year following the DSA's date of application, as defined by the Commission, "and in any event prior to deploying functionalities that are likely to have a critical impact on the risks identified pursuant to this Article." At this time, there is no available interpretation of "critical impact," making it difficult to identify events that justify a new assessment.

We suggest factors that could meet the minimal threshold of "critical impact" given the circumstances at play (non-exhaustive list):

- **Substantial changes to VLOP or VLOSEs' terms and services (beyond purely linguistic edits, e.g., new content policies or enforcement guidelines; change of ownership), which could exacerbate the disproportionate removal of content, especially if removal is automated;**
- **A new product or feature that may significantly impact fundamental rights, either by increasing the severity or scale of existing negative impacts, or by impacting a new fundamental right or new group of people, in a new way;**
- **Exceptional external circumstances and/or crises, such as conflict or war, geopolitical tensions, emergency measures, security threats, climate disasters, and/or key elections.**

Many open questions remain, however. First, what should be the baseline for determining "regular" systemic risks, and would anything above that baseline be considered irregular or exceptional? Second, should any additional FRIA be conducted as extensively as the annual one, or can the process be shorter and more targeted? Third, how does this additional FRIA fit into VLOPs and VLOSEs' overall due diligence obligations in the DSA?

In terms of process, rather than an entire ad hoc procedure, we recommend that VLOPs and VLOSEs conduct ongoing due diligence throughout their products and systems' entire lifecycle. They should conduct "exceptional" FRIAs whenever there is a heightened risk to fundamental rights, especially those of vulnerable recipients of the services, and when assessing and taking measures to mitigate these risks could help reduce harm.

We recognize that conducting meaningful FRIAs takes time and resources, as well input and participation from external stakeholders who should participate in the process, and we acknowledge accordingly that, at times, other harm prevention processes (e.g., red teaming, sandboxing, targeted consultations with external stakeholders through "Trust and Safety Advisory Councils'," etc.) may be more adequate.

2.5. How transparent should the assessment be?

The transparency of FRIAs enables external stakeholders, including civil society organisations, researchers, journalists, and people impacted by systemic risks, **to scrutinise the impact assessment and ensure it is more than merely a "tick box" exercise**. Transparency is crucial for explaining how exactly the FRIA may have influenced VLOPs and VLOSEs' design and development of their services, including algorithmic systems.

As per Article 42(4) of the DSA, VLOPs and VLOSEs must publish extensive documentation of the risk assessment and auditing process, including a report on the results of the risk assessment, the specific mitigation measures implemented, and audit documentation, including the audit report and implementation report. While the DSA does not explicitly require the publication of the entire risk assessment, including the FRIA, Recital 100 of the DSA emphasises the importance of comprehensively reporting on risk assessments, given the heightened risks related to the functioning of VLOPs and VLOSEs.[26]

Specifically, we expect the report to include, as a minimum, the following elements for FRIAs:

- **A detailed methodology of the FRIA**, including an explanation of how the assessment was carried out and which systems and functionalities were assessed, and whether any exceptional assessments were conducted in addition to the annual one;

26 In view of the additional risks relating to their activities and their additional obligations under this Regulation, additional transparency requirements should apply specifically to very large online platforms and very large online search engines, notably to report comprehensively on the risk assessments performed and subsequent measures adopted as provided by this Regulation.

- **A detailed mapping and description of the service and its overall structure**, including for any algorithmic systems, so the public can understand the platform's architecture, its functioning, and the interdependence between its different systems or elements;

- **The purpose and key design specifications of each assessed algorithmic system**,[27] including the algorithm's general logic, the rationale and assumptions made with regard to potentially-affected persons or groups of persons, the main classification choices, what the system is designed to optimise for and the relevance of the different parameters, and the decisions about any possible trade-offs;[28]

- **Identified negative effects for the six fundamental rights clusters**, an explanation of which negative effects were assessed as the most significant based on their severity and probability, a classification of the negative effects, and a clear indication of which part of the service they stem from;

- Measures or actions already taken, or to be taken, following the impact assessment, including:

    - Any **mitigation measures** adopted for each negative effect, as per Article 42(4)(b), including those adopted following the audit,

    - Any negative effects assessed by the platform or indicated by auditors as being **unacceptable under the Charter**, including an explanation of steps taken by the platform to prevent the unacceptable negative effect, e.g. changing the design of the service, discontinuing a specific system or policy, or refraining from introducing a new system or policy;

- **Which internal departments, teams, experts** were involved in the FRIA **internally**, including senior/executive leaders responsible for approving the assessment and implementing the audit;

- **Which external stakeholders were consulted** as part of the FRIA, when and how they were engaged with, and any **outcomes of these consultations**, including any details of how the VLOP or VLOSE responded to external stakeholders' input and integrated it in their services and activities.

Under Article 42 of the DSA, VLOPs and VLOSEs must publish this information within a maximum of three months after they receive an audit report from the auditor. However, negative effects on fundamental rights can only be comprehensively and accurately identified **if civil society can review and meaningfully contribute to impact assessments before they are audited**.

As a best practice, we recommend that the European Commission provides civil society organisations with access to the full risk assessment, including the FRIA and all relevant supporting documentation, at an early stage, before the assessment is audited and before the summary is published. This could be done under conditions of confidentiality and/or could be limited to members of a civil society advisory group, which we recommend being established.[29] This would allow civil society to contribute meaningfully to the process and support both the Commission and VLOPs/VLOSEs in accurately identifying negative effects on fundamental rights, ultimately helping them to prevent any fundamental rights violations.

27 Trade secrets should be interpreted in line with EU law, i.e. be limited to information that has commercial value. In this context, platforms cannot use the argument of trade secrets to hide key information about the negative effects that their services have on fundamental rights.

28 This wording is inspired by Annex IV (2)(b) of the proposal for the Artificial Intelligence Act.

29 See the proposal put forward by Dr Suzanne Vergnolle in "Putting collective intelligence to the enforcement of the Digital Services Act" https://www.article19.org/resources/report-civil-society-enforce-eu-digital-services-act/

3. STAKEHOLDER INVOLVEMENT IN THE FRIA PROCESS

Meaningful stakeholder engagement is critical at every stage of FRIAs. We recommend involving a broad and diverse set of actors in the FRIA process, including those who can influence or are influenced by the design, functioning, or use of VLOPs and VLOSEs' services and related systems, including algorithmic systems.

A wide range of individuals, groups, and institutions should include those working internally for the VLOP/VLOSE as staff, contractors, or as part of companies' value chains ("workers"), as well as external individuals and groups such as rights-holders, affected communities, civil society organisations, academics, and other experts ("external stakeholders").

VLOPs and VLOSEs should implement the following recommendations, and the European Commission should verify their adequate implementation accordingly.

We refer to "stakeholder mapping" as the process of identifying who should be part of the process. When mapping potential external stakeholders, we recommend that VLOPs and VLOSEs focus first on those most impacted by their systems that are designed and used, especially historically and institutionally marginalised groups, and then work outwards. Centering at-risk groups helps with better understanding the problem and with prioritising negative effects on fundamental rights based on the severity and likelihood of harm.

3.1. What are VLOPs and VLOSEs workers' roles and responsibilities in the FRIA process?

Regardless of who from the VLOP/VLOSE is actually tasked with carrying out the FRIA internally (known as the "assessor"), the process should involve representatives from cross-functional (XFN) teams (i.e. teams across internal departments) to capture all aspects of fundamental rights and to avoid any aspects becoming siloed.

It is particularly relevant to involve teams of people working in engineering, product development, research, risk management, legal, policy, finance, sustainability, communications, marketing, sales, human resources, trust and safety, and human rights. There is also a role to play for product managers and workers who coordinate various teams internally.

Moreover, fundamental rights experts should either be actively involved in assessing risks, or should inform the risk assessment process. Once internal working groups are identified and convened, they should further map which external stakeholders may be important to engage with, and should continue to play an important role throughout the FRIA process, particularly when it comes to communicating externally on progress and findings.[30]

Finally, Article 41 of the DSA requires that VLOPs and VLOSEs establish a "compliance function" with "compliance officers" responsible for, among other things, "ensuring that all risks referred to in Article 34 are identified and properly reported on and that reasonable, proportionate and effective risk-mitigation measures are taken pursuant to Article 35" (Article 41(3)b DSA). The VLOP/VLOSE's management body must also be involved in FRIAs, with the mandate to monitor and approve FRIA-related strategies and policies on at least an annual basis.

30 In the context of the EU AI Act, which is relevant here, Kalvi and Kotsinos note the importance of "meaningful communication between the various expertise involved in an AIA process and the public, considering that, as demonstrated in the context of EIA, a one-size-fits-all approach to providing information may prevent non-experts from expressing their opinions about the initiatives under assessment." <https://dl.acm.org/doi/10.1145/3593013.3594076>

We recommend that the European Commission carefully review which members of senior management, leadership, and the executive team are involved in approving and supporting any mitigation measures resulting from the FRIA, as required under Article 35 of the DSA. This can help ensure that the FRIA process actually leads to the meaningful implementation of risk mitigation measures.

3.2. What is the role of external stakeholders in the FRIA process?

In the context of FRIAs, we understand external stakeholder engagement to be listening and collaborative processes, which allow individuals outside of the VLOP/VLOSE influenced by or who can influence their services to meaningfully inform how negative effects on fundamental rights are identified, analysed, and assessed under Article 34(1)b of the DSA. Their insights should also inform how to prioritise and assess the severity and probability of negative effects, and how to mitigate them in line with Article 35 of the DSA.

We acknowledge that VLOPs and VLOSEs' ability to engage with external stakeholders depends on their capacities and resources. While every designated VLOP/VLOSE should foster meaningful stakeholder engagement, they are not a monolith; some are, for instance, non-profit and community-governed projects that mainly serve the public interest.[31] Accordingly, the following recommendations are not intended as a "one size fits all" approach. Ultimately, how VLOPs and VLOSEs engage with stakeholders should vary depending on the scale and severity of fundamental rights risks, their platform model (e.g. for profit vs public interest/ user-based vs community-oriented), and their available resources (since larger companies tend to have more capacity and resources to engage more regularly and with a broader range of stakeholders).

Engagement is particularly important when assessors and VLOPs/VLOSEs need to make difficult decisions, such as deciding whether to strengthen privacy or increase transparency, and particularly when such decisions have serious implications for marginalised groups or wider society. By engaging with affected communities on their priorities, it often becomes clear that different fundamental rights are interdependent and complementary, rather than being in conflict or competition.

While the DSA does not mandate external stakeholder engagement, Recital 90 of the DSA clearly establishes the responsibility for VLOPs and VLOSEs to do so.[32] We can also interpret Article 42(4)(e) of the DSA, which requires VLOPs and VLOSEs to provide "information about the consultations conducted by the VLOP/VLOSE in support of the risk assessments and design of the risk mitigation measures," as establishing a duty for platforms to consult with external stakeholders.[33] As recommended by ARTICLE 19 in their 2023 report on collaborations between the European Commission and civil society, expert groups as mandated under Article 3 of the Commission Decision,[34] can also help with the risk assessment and engagement processes, including carrying out meaningful public consultations.

31 Jan Gerlach, The Digital Services Act could require big changes to digital platforms. Here are 4 things lawmakers need to know to protect people-powered spaces like Wikipedia (2021).

32 Under Recital 90 of the DSA, providers "should ensure that their approach to risk assessment and mitigation is based on the best available information and scientific insights and that they test their assumptions with the groups most impacted by the risks and the measures they take. To this end, they should, where appropriate, conduct their risk assessments and design their risk mitigation measures with the involvement of representatives of the recipients of the service, representatives of groups potentially impacted by their services, independent experts and civil society organisations. They should seek to embed such consultations into their methodologies for assessing the risks and designing mitigation measures, including, as appropriate, surveys, focus groups, round tables, and other consultation and design methods."

33 Note that in the spirit of ensuring the integrity of the EU regulatory landscape we borrow some of the terminology from the proposed directive on Corporate Sustainability Due Diligence (CSDDD), which many VLOPs and VLOSEs might be subjected to (in the Parliament version). At the moment of writing, CSDDD requires that "[c]ompanies shall also carry out meaningful engagement (...) with potentially affected stakeholders including workers and other relevant stakeholders to gather information on as well as to identify and assess actual or potential adverse impacts" (Article 6, paragraph 4 CSDDD). These responsibilities build on Principle 18 of the UNGPs, which requires companies to "(a) [d]raw on internal and/or independent external human rights expertise; [and] (b) [i]nvolve meaningful consultation with potentially affected groups and other relevant stakeholders, as appropriate to the size of the business enterprise and the nature and context of the operation" when conducting human rights impact assessments.

34 Commission Decision establishing horizontal rules on the creation and operation of Commission expert groups, C(2016) 3301, 30 May 2016, https://ec.europa.eu/transparency/documents-register/detail?ref=C(2016)3301&lang=fr

3.3. How often should consultations with external stakeholders take place?

External stakeholder engagement, as part of FRIA processes, should happen at least annually, but we encourage more frequent and regular engagement as needed. Where VLOPs and VLOSEs have pre-existing relationships with external stakeholders, such as civil society groups and affected communities, it is worth involving them from the start of the FRIA process, when defining the purpose and desired outcomes of the assessment, as well as during the initial stakeholder mapping.

Under Article 34(1) of the DSA, risk assessments should be carried out at least annually and "prior to deploying functionalities that are likely to have a critical impact on the risks identified." Indeed, while engagement is often seen as a one-off event, it generally works best as a dynamic, iterative process with several objectives, involving diverse target groups, and using varying methods depending on the timing, purpose, and objectives of a given FRIA process. In any case, engagement should happen where contributions can be most influential for the FRIA, especially marginalised groups and those with lived experience. Such groups' assistance can also be invaluable for further stakeholder mapping, as they can suggest and connect with other relevant groups in order to involve those who are not typically consulted or who lack strong existing networks.

3.4. What are the criteria for meaningful external stakeholder engagement?

Building on ECNL's Framework for Meaningful Engagement informed by over 180 consultations, we outline three core conditions that must be satisfied for engagement in the FRIA process to be considered meaningful:

1. SHARED PURPOSE

The engagement must have a purpose and desired outcomes beyond the VLOP/VLOSE's self-interest or desire to merely fulfil compliance obligations under Article 34(1)(b) of the DSA. Its aim should be to advance either the specific interests of potentially affected people or the overall public interest purpose of assessing and mitigating risks.

2. TRUSTWORTHY PROCESS

The engagement process must be inclusive, open, fair, respectful, and delivered with integrity and competence. The VLOP or VLOSE must be open and honest about any limitations or barriers to conducting the engagement or incorporating stakeholders' insights.

3. VISIBLE IMPACT

External stakeholders must have sufficient influence on identifying, analysing, and assessing the systemic risks of VLOPs and VLOSEs' services, with the ultimate goal of determining risk mitigation measures under Article 35 of the DSA. The VLOP/VLOSE must be open about how any trade-offs or competing priorities may result in the FRIA (and any resulting future risk mitigation measures) diverging from participants' expectations.

4. STAKEHOLDER ENGAGEMENT BENCHMARKS

HYPOTHESIS 1

Despite being best placed to identify actual or foreseeable negative effects of VLOPs and VLOSEs' services on fundamental rights, affected communities and civil society organisations are too often excluded from the FRIA process.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- Which substantive issues of the FRIA process do external stakeholders engage with?
- Do assessors engage external stakeholders in the process of evaluating negative effects on fundamental rights stemming from the design and use of all services, especially machine learning or algorithmic-based systems for content governance (such as content moderation and recommender systems), targeted advertising, and data practices?
- What possibilities do external stakeholders have to inform the assessment of negative effects' severity and probability? Do they have the opportunity to input on whether, given these conditions, it is appropriate to design or use the service(s), or to weigh in on the mitigation measures needed to address these negative effects, as required by Article 35 of the DSA?

HYPOTHESIS 2

VLOPs and VLOSEs rarely consult with external stakeholders who have fundamental rights expertise and/or lived experience, thus failing to ensure that their services are shaped by the concerns and rights of those most affected, often already marginalised, groups.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- How do assessors map stakeholders and ensure that the process includes a wide range of voices, particularly those who may be most negatively impacted such as historically and institutionally marginalised groups, and those with lived experience?
- How does the process incorporate historically underrepresented, ignored or hard-to-reach groups' needs?
- When it is not possible to engage directly with affected communities, do the assessors consult instead with proxy groups (i.e. people who can speak about the experiences of a particular community without necessarily belonging to the community in question) such as civil society organisations, activists, academics, or researchers? How do assessors map these groups and ensure they can provide affected communities' viewpoints from the start of the FRIA process?

HYPOTHESIS 3

For any number of historic reasons, including previous "participation washing" where any external stakeholder engagement was purely performative, external stakeholders may distrust a specific VLOP/VLOSE, or an entire sector.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- How do assessors account for issues of power dynamics and stakeholder accessibility?
- How do assessors communicate about the intended purpose and desired outcomes of stakeholder engagement?
- How can stakeholders' perspectives best influence the FRIA, and when in the process are they actually involved?
- How do assessors show openness to identifying and analysing all negative effects on fundamental rights based on the scope of Article 34(1) (b) of the DSA, rather than just a selected few risks?
- How are participants kept informed about how their feedback is included or how it influenced the FRIA, and about any resulting mitigation measures? How are they updated on specific trade-offs and competing priorities that may have overridden their feedback?

HYPOTHESIS 4

Stakeholder engagement is too often a one-way-street, with platforms seeking input from external stakeholders, but failing to share their own internal information.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- What kind of information is shared with consulted stakeholders?
- How do assessors ensure that consulted stakeholders have adequate knowledge and/or understanding of their services, systems, products, and policies, in order to foster constructive, informed feedback?
- How do platforms share internal information with external stakeholders, and support those stakeholders' own interests as part of the FRIA process, so that the engagement isn't solely extractive?

HYPOTHESIS 5

Each engagement process has its own barriers and limitations, some of which may not be obvious or even foreseeable. These may include constraints on the overall purpose or outcomes of the engagement or the FRIA, as well as constraints on funding, resources, capacity, competence, knowledge, or expertise, among others.

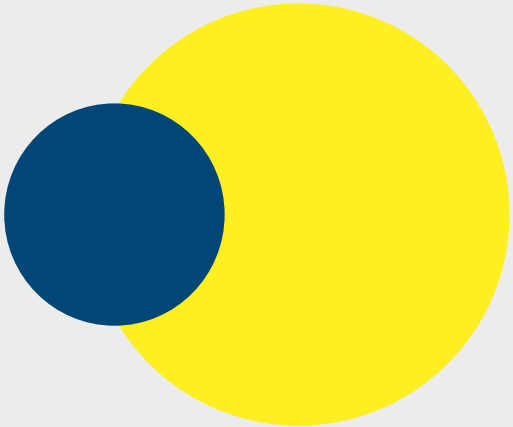
EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- How early in the process do assessors identify constraints to meaningful engagement? Do assessors document these barriers and any attempts to overcome them, and make this information publicly available?
- When in the day, and in which format, does the engagement take place? How do assessors ensure that the timing and format is convenient for consulted groups?

- How is external stakeholders' participation supported and resourced, e.g., with financial compensation for stakeholders' time, travel expenses, etc?
- How, and by whom, is the engagement methodology designed? How does the convenor ensure that the engagement is culturally sensitive, participatory, and relevant, especially for non-experts?
- What language support is made available? How do assessors ensure any terminology used is broadly appropriate and understandable?
- How do assessors ensure participants' safety and security, especially for people most at risk of harm?

PART II. ESTABLISHING BENCHMARKS FOR ASSESSING THE NEGATIVE EFFECTS OF AUTOMATED CONTENT MODERATION ON FREEDOM OF EXPRESSION

In the second half of this paper, we focus specifically on content policy and automated content moderation systems, based on a number of hypotheses related to the potentially negative effects of VLOPs and VLOSEs'' content policies and systems on fundamental rights, especially the right to freedom of expression and information. With each hypothesis, we provide the European Commission with suggested evaluation benchmarks, so they can ask VLOPs and VLOSEs the right questions and effectively evaluate FRIAs. While these hypotheses and benchmarks are non-exhaustive, they convey the level of detail we would expect the Commission to require from assessments.



1. CONTENT POLICIES

HYPOTHESIS 1

The process of developing content policies is designed to provide VLOPs and VLOSEs with a large margin of discretion. Vague wording can have negative effects on fundamental rights and may disproportionately impact vulnerable service recipients. Shortsighted and opaque content policy development can also lead to discrimination.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- Are VLOPs or VLOSEs' content policies and internal content moderation guidelines structured per the type of alleged illegal or potentially harmful content that they seek to combat? Examples include:
    - Manifestly illegal content irrespective of its context, such as child sexual abuse material;
    - Alleged illegal content, including terrorist or violent extremist content online, incitement to discrimination, hostility, and violence;
    - Alleged illegal or legal forms of hate speech;
    - Sexual harassment or other discriminatory action;
    - Content which perpetuates gender-based online violence;
    - Non-consensual nudity;
    - Misinformation and disinformation.

- When assessing a qualitative description of content policies, including its precise purposes, how do VLOPs and VLOSEs incorporate protective safeguards to prevent discriminatory bias towards vulnerable recipients of the service? Concretely, do content policies adequately consider and address the following:
    - The open-ended list of grounds protected against discrimination as prescribed by Article 21 of the EU Charter;
    - Specific languages, including regional and minority languages in EU Member States, as defined in the European Charter for Regional and Minority Languages;
    - Reclaimed language, i.e. language that used to be slurs against vulnerable recipients of the service, but that now has a pro-social function used by these groups to cope with hostility and/or identify themselves; and
    - The potential negative impact and collective harm of their content policies on specific entire groups consisting of vulnerable recipients of the service, including national and regional minorities, persons with migrant or refugee status, gender identity or sexual orientation, journalists, human rights activists etc.?

- Do VLOPs and VLOSEs conduct a gap analysis of their content policies to ensure their alignment with the EU Charter, and if so, how frequently? What factors, including geopolitical changes in EU Member States, trigger such a gap analysis? When conducting a gap analysis, what sources are considered and incorporated into the assessment of current content policies? For instance, do VLOPs regularly consult the European Union Minorities and Discrimination Survey?

- How do VLOPs ensure that their content policies comply with Article 25 of the DSA, i.e. ensuring that content policies are designed, organised, and enforced in a manner that enables recipients of the service to make informed decisions about what content they are allowed to share and to understand the consequences of infringing VLOPs' and VLOSEs content policies?

HYPOTHESIS 2

Inconsistent, uneven, or biased enforcement of content policies and internal content moderation guidelines can exacerbate existing societal discrimination, power imbalances, and inequalities. Over- and underenforcement of content policies (i.e. false positives, where legitimate content is removed, or false negatives, where prohibited content remains) disproportionately impacts, and may indirectly silence, vulnerable recipients of the service.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- When assessing content for immediate removal or other forms of content restrictions, what thresholds or tiered approaches, if any, are included in content policies and internal content moderation guidelines, and how are these defined? Examples of such thresholds used by VLOPs and VLOSEs include the threat of "imminent physical harm" or "real world harm."
- Do VLOPs and VLOSEs adequately assess whether their thresholds for imminent online content removal align with international human rights law, including Article 20 of the ICCPR and the Rabat Plan of Action?
- Are enforcement standards explained to recipients of the service in unambiguous and plain language, as prescribed by Article 12 of the DSA?
- How are EU Member States' contextual diversity and demographics reflected in the enforcement of content policies? Do VLOPs and VLOSEs consider local context when developing content policy and enforcement guidelines, especially forregions or Member States at risk of ongoing and systematic violations of fundamental rights and weakened rule of law?
- Are content policies translated, available, and easily accessible in all EU Member States' national, regional, and minority languages?
- Do VLOPs and VLOSEs have a list of identified existing, potential, and future negative effects on fundamental rights for each EU Member State? Are these lists made available to representatives of the recipients of the service, representatives of groups potentially impacted by their services, independent experts, and civil society organisations? Are these stakeholders consulted throughout the process of developing these lists?

HYPOTHESIS 3

Most VLOPs and VLOSEs publish community guidelines that provide explanations to recipients of their service on how they govern user-generated content, yet their comprehensive content moderation policies remain opaque. Newly-formed, reactive rules are adopted as new challenges, such as the escalation of a conflict, emerge, yet these are scattered across platforms and announced in multiple blog posts and press releases, including third party websites or corporate profiles. Policy changes and amendments are incorporated incoherently into existing policies, remaining unclear and vague, without providing any systemic solutions to issues at stake.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- Are VLOPs' and VLOSEs internal content moderation guidelines made partially or fully available to the European Commission and to the Board of Digital Service Coordinators, in order to adequately assess their compliance with the EU Charter? If not, what reasoning is provided to justify this lack of transparency?
- What interface design tools enable direct and easy access to their service's content policies?
- Where on the platform are content moderation policies published? Are VLOPs and VLOSEs' content policies concentrated in one place, which is easily accessible to recipients of the service?

- How does the service communicate about any carve-outs or exceptions to existing content policies and internal content moderation guidelines, or other extraordinary measures, made public on the service? Are these announced in unambiguous language? Are they specific, predictable, and limited in time (sunset clauses)?
- Are they proactively communicated in all EU Member States' national, regional, and minority languages, with appropriate enforcement justification that complies with the EU Charter?

HYPOTHESIS 4

VLOPs and VLOSEs' business interests impact and inform their content moderation policies and their enforcement, in ways that may exacerbate negative effects on the rights to freedom of expression and opinion, non-discrimination, and other fundamental rights.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- How do VLOPs' and VLOSEs 'advertising policies or other business interests influence content policies and their enforcement?
- Do VLOPs and VLOSEs have an ads transparency centre, and if so, what type of information do they include?
- Are some types of content removed or demoted, if it could impact advertisers' willingness to purchase advertising on the platform? How has this been accounted for in the assessment?
- How do VLOPs and VLOSEs prevent their monetisation programs from channelling income to actors associated with sanctioned entities, or to foreign and local actors systematically producing and/or distributing disinformation and other types of alleged illegal or potentially harmful but legal online

  content?

2. ALGORITHM DESIGN AND USE

HYPOTHESIS 1

The upload filters commonly used by platforms constitute prior restraint on content, which is a violation of users' rights to freedom of expression, the protection of personal data, and the confidentiality of communication. [35]

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- Are algorithms used in a way that constitutes prior restraint (e.g. through upload filters), or only to flag notices of illegal or violative content, or both?
- What objectives do VLOPs and VLOSEs pursue when using upload filters*[36]?
- Is each specific algorithm effective and necessary to achieve its objective*?
- What other content moderation measures were considered and why were they discarded*?
- Where the VLOP or VLOSE identified an alternative to upload filters, which was less effective but appeared to have fewer negative fundamental rights effects, what competing interests justified discarding the alternative*?
- Do the upload filters' objectives justify the negative fundamental rights effects? Did the VLOP or VLOSE conduct a balancing test to ensure that the negative effects on fundamental rights are proportionate?*
- How did the VLOP or VLOSE address the negative effects on the right to confidentiality of communication resulting from the use of upload filters?

HYPOTHESIS 2

Lists of allowed and/or prohibited words included in automated content moderation systems and internal content moderation guidelines (so-called "allow and deny lists") can be helpful, but also imperfect and inconsistent, in moderating content. A lack of transparency around how such lists are compiled may reinforce existing discriminatory biases and lead to under- and overenforcement.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- If platforms use "allow and deny lists," what do these include, how are they created, and how often are they updated?
- How do platforms ensure that these lists do not discriminate against vulnerable groups?
- How is linguistic diversity accounted for in the creation of these lists?

35 Paragraphs 48-50 https://curia.europa.eu/juris/document/document.jsf?docid=119512&doclang=EN

36 The questions with an asterisk have been inspired by the FRIA methodology developed in the Netherlands for algorithms used in the public sector: https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms

HYPOTHESIS 3

Datasets (including training datasets) and algorithms for automated content moderation systems can cause or contribute to negative effects on fundamental rights, especially the right to non-discrimination.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- What training methodologies and techniques and training datasets do VLOPs and VLOSEs use to develop and deploy algorithmic content moderation systems?
- What are the sources, scope, and main characteristics of training, validation, and testing datasets?
- How was the data obtained, selected, labelled, and cleaned?
- Are these datasets relevant and representative in light of the algorithm's objective, free of errors, and complete?
- Do VLOPs and VLOSEs define bias and discrimination in accordance with the EU Charter and the open-ended list of protected characteristics?
- What safeguards are implemented by VLOPs and VLOSEs to ensure data quality and prevent bias and discrimination?
- Do these datasets consider characteristics or elements that are specific to the geographical, behavioural, or functional setting within which the content moderation system is deployed? Does data quality differ depending on context, e.g., geographic region, language, or type of content? If so, how? Is the difference in data quality proportionate given the negative effects to fundamental rights, and why?

HYPOTHESIS 4

The algorithm's performance can reinforce existing bias and discrimination, as well as patterns of systemic oppression.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- How do VLOPs and VLOSEs validate and test algorithms?
- How often are algorithmic content moderation systems (including models) updated? Are they updated following a regular cadence, or only following specific trigger events, and if so, which events are these?
- When developing algorithmic content moderation systems, what level of accuracy do VLOPs and VLOPs accept? What metrics are used to define it? Did the VLOP or VLOSE conduct a balancing test to ensure that the negative effects of errors and loss functions are proportionate?
- How is performance measured (i.e. what is the definition of success and/or optimisation goal)? What definitions of fairness are used?
- How is performance measured (i.e. what is the definition of success and/or optimisation goal)? What definitions of fairness are used?
- How many users are affected by false positives (i.e. over-enforcement of content policies) and, if this information is available, false negatives (i.e. under-enforcement)?
- Does the overall use of the algorithmic system result in a reasonable balance between the objectives pursued and the fundamental rights that may be infringed? Who makes this decision? Are people with fundamental rights expertise involved in the decision-making process?

- Do VLOPs and VLOSEs regularly verify that algorithms are functioning as intended? Are algorithms adjusted to mitigate biases as they are discovered and to ensure fairness?

HYPOTHESIS 5

The platform accepts different standards for algorithms' performance, based on the varying level of harm caused by different types of content or contexts, for example:

- Illegal content including terrorist or violent extremist content, child sexual abuse material, illicit and regulated good and services;
- Abuse, harassment, doxxing, gore or violent content;
- Misinformation and disinformation;
- Impersonification, spam;
- Specific languages or regions; and
- Specific groups, e.g. activists, political dissidents, journalists, and historically or institutionally vulnerable recipients of the service.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- Do VLOPs and VLOSE accept lower accuracy for specific types of content or is the algorithmic system designed to perform equally across all content? If the former, for which type of content is different performance accepted, and why?
- How does the level of accuracy differ between languages, geographic regions, or types of users (e.g. public figures)? How do VLOPs and VLOSEs justify any discrepancies, and assess the impacts of these discrepancies on fundamental rights?

HYPOTHESIS 6

Errors in automated moderation (false positives and negatives) are increased and exacerbated by inadequate or non-existent human review and/or rare or irregular system updates, thus putting vulnerable recipients of the service at risk.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- Where content is automatically flagged, removed, or restricted from view, who verifies whether the decision was correct or erroneous, and how do they make that assessment? What is the process for review and remedy? Which teams are involved in the process?
- Are users informed when their content is removed or restricted by an algorithmic system? How do VLOPs or VLOSEs ensure that users have effective access to remedy? What is the process and its stages?
- What percentage of cases are reviewed first by humans? What percentage of automated decisions are subsequently reversed by humans? Is there a breakdown of accuracy per demographic group, particularly those at risk? Indicate the percentage of reversals following an appeal by a user and other situations.
- How is automation bias of human reviewers accounted for and prevented?

HYPOTHESIS 7

Profiling users for the purposes of content moderation (e.g. to predict which users are most likely to post illegal content) can violate the rights to freedom of expression, data protection, and privacy.

EVALUATION BENCHMARKS FOR THE EUROPEAN COMMISSION

- Do VLOPs or VLOSEs profile users algorithmically, to predict who is likely to post illegal or violative content?
- If so, how do they comply with legal obligations around data protection? How do they protect fundamental rights (e.g. the right to information on processing, the right to object)?
- What happens when a user is identified as presenting a high risk of posting illegal content or content that violates terms and conditions? Does the user have any way to contest this designation?
- Do the VLOPs and VLOSEs recognise this process as constituting automated decision-making under Article 22 of the GDPR, and if so, why? If so, how do they comply with additional rights and safeguards set out in Article 22?

CONCLUSION

For years, civil society organisations have been advocating for human rights-based regulation of Big Tech. Now, for the first time in the history of internet regulation, VLOPs and VLOSEs are legally obliged to conduct FRIAs. Their compliance will be monitored by the European Commission, the guardian of EU law with strong executive powers, which must ensure that this novel tool is not downplayed by private actors seeking to evade regulatory scrutiny. Enabling, achieving, and safeguarding meaningful, effective FRIAs will be essential for achieving this goal.

Enforcing the protection of fundamental rights and freedoms guarantees personal freedom and dignified existence for all. Yet, the societal risks and human rights abuses caused by VLOPs and VLOSEs' services tend to be collective in nature, going beyond violations of individual rights. For instance, addressing fundamental rights risks from automated content moderation at scale via access to individual remedy is challenging. Mandatory FRIAs in the DSA will ensure that the safeguarding of fundamental rights is built into VLOPs and VLOSEs' systems, processes, and policies, while also creating an essential layer of protection that could deliver collective remedy, especially to vulnerable recipients of the service.

All stakeholders have a role to play in making the DSA, and more concretely FRIAs, a success. This includes civil society organisations, academics, national Digital Service Coordinators, the European Board for Digital Services, and the European Commission. The European Commission's newly-established DSA enforcement team will need to operationalise the new enforcement model for ensuring Big Tech companies fulfil their obligations – a challenging task. Yet done right, it could transform the online ecosystem to finally serve people, and not corporate profit, first.

Access Now, ECNL, and our partners look forward to working with the European Commission and the co-legislators to strengthen DSA enforcement and effectively guarantee the protection of fundamental rights and freedoms for all.

accessnow

Access Now Europe
Rue Belliard 12
1040, Brussels
Belgium
www.accessnow.org
@accessnow

European Center for
Not-for-Profit Law

European Center for Not-for-Profit Law
Riviervismarkt 5
2513 AM, The Hague
Netherlands
www.ecnl.org
@enablingNGOlaw