Search for
Common Ground
Trust, Collaboration, Breakthroughs

Boosting the EU Digital Services Act:

Conflict Sensitivity Requirement for Very Large Online Platforms

Christian Cirhigiri, November 2023

About

The Digital Services Act (DSA) is a binding European Union(EU) legislation that regulates all hosting services and online platforms that offer services within the Union. The legislation is enforced by relevant authorities in EU member states and the European Commission, with the latter holding a supervisory role over its implementation by all Very Large Online Platforms and Search Engines. It sets the rules for tackling illegal content as well as aims to address various challenges posed by the digital environment, such as harmful online content, misinformation, and the power of dominant online platforms. All online intermediaries in scope should be preparing for full compliance in February 2024.

The report examines how a "conflict sensitivity" approach can improve future risk assessments of Very Large Online Platforms (VLOPs) as stipulated in Article 34 of the DSA, which requires platforms to "identify, analyze 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." We highlight lessons learned from the first round of risk assessments of VLOPs and examine digital peacebuilding case studies from Sri Lanka, Kenya, and the Democratic Republic of Congo.

This policy report does not propose a new parallel compliance mechanism. Our recommendations are intended to help the European Commission, the Directorate General -Communications Networks, Content and Technology (DG-CNECT), and the newly established DSA enforcement team to play a preventive role in addressing online manifestations of conflicts in the Union before they escalate into offline violence, thereby also setting a standard for the rest of the world.

This report is realized with the support of the Belgian development cooperation. The opinions expressed by our organization do not necessarily reflect those of the Belgian State and do not bind the latter in any way.

Acknowledgments

This policy report is produced by Search for Common Ground (Search), a global peacebuilding organization headquartered in Brussels and Washington DC, with a 42-year track record of supporting local actors to find local solutions to today's most brutal, violent conflicts. Currently working in 36 countries, Search for Common Ground has applied conflict sensitivity to tackling online manifestations of conflicts in countries like Kenya, the US, Nigeria, Sri Lanka, the Democratic Republic of Congo, and others, where electoral violence and political upheaval linked to online polarization have negatively impacted social cohesion.

There are many colleagues who contributed to this report, including Mike Jobbins, Lena Slachmuijlder, Charline Burton, Inge Brees, Katie Smith, Laurent Kasindi, Durga Pulendran, Dominick Mwambui, Adrienne Lemon, Jack Farrell. Minji Kim assisted with this report's design. Special thanks to Asha Allen, Advocacy Director at the Centre for Democracy and Technology, Europe Office and Ravi Iyer, Managing Director of the Psychology of Technology Institute, for the helpful review of the report.

To Cite this report

Christian Cirhigiri. "Boosting the EU Digital Services Act: Conflict Sensitivity Requirement for Very Large Online Platforms." Search for Common Ground (November 2023).

Table of Contents

Ce sont des militaires kényans engagés contre les terroristes, et non des M23 capturés dans le véhicule de la Monusco en RDC

Executive summary

The internet has brought numerous benefits to society, but it has also posed unforeseen challenges. With Very Large Online Platforms (VLOPs) now an integral part of our daily lives, these platforms collect vast amounts of user data and have a significant influence over online interactions. To address this, the European Union (EU) acknowledges the importance of regulating VLOPs to ensure the protection of human rights and promote a safe online environment. However, the most recent legislative effort, the Digital Services Act (DSA), lacks clarity in critical areas such as risk assessment, transparency, and stakeholder engagement, posing challenges to its effective implementation.

What has not been looked into yet, is the impact that the DSA and its attempt to regulate tech companies behavior could have on online/offline violence in the rest of the world, including in conflict-affected countries, and conflict-affected and fragile settings, nor which lessons can be learnt from the rest of the world to improve the implementation of the DSA in Europe and beyond. This report seeks to address these gaps.

The report is divided into five sections. The first section explains what conflict-sensitivity is and why it is important for the European Commission. The second section examines how three major social media platforms - Facebook, Snapchat and TikTok - have dealt with conflict sensitivity, based on their first risk assessments. The third section shares valuable insights to address systemic risks on VLOPs in Kenya, Sri Lanka, and the Democratic Republic of Congo, which can improve the DSA's implementation. The fourth section highlights the importance of involving non-EU actors in implementing the DSA, and the fifth section summarizes the report's key recommendations.

Recommendations

- The newly established DSA enforcement team of the European Commission should include conflict sensitivity as a core requirement into risk assessment guidelines

- The newly established DSA enforcement team of the European Commission should enable a multi-stakeholder process (i.e with civil society, industry, and EU input) to formulate risk-assessment guidelines and ensure a meaningful participation of conflict-affected communities in developing these guidelines.

- The European Commission, and in particular DG-CNECT, should engage and sustain a policy dialogue with CSOs in conflict-affected countries on the DSA by conducting quarterly public consultations with relevant authorities and CSOs in these countries on addressing systemic risks on VLOPs

- The European Commission, and in particular DG-CNECT, should require platforms to publicly provide product experimentation results on outcomes of societal interest for any meaningful product design decision.

- The European Commission should ensure that the impacts of recommender systems, as a crucial design layer of large social media platforms, are assessed comprehensively from a conflict-sensitive and human-rights perspective.

Digital Peacebuilders Guide
Search for Common Ground
BUILD UP
ConnexUs

Introduction

“Technology... is a queer thing; it brings great gifts with one hand and stabs you in the back with the other.” Charles Percey Snow, English Novelist.

In a matter of years, internet users have transitioned from simply searching on Google to depending on it for multiple tasks such as directions, calendars, address books, entertainment, relationship advice, voice mail, and telephone calls. From a simple social networking tool started in the 2000s, Facebook currently hosts over 3.9 billion active users who post content, like, share or comment on their friends’ updates. It is increasingly hard to do anything online these days without using Google, Facebook, or Amazon, and Very Large Online Platforms (VLOPs) collect vast amounts of user data that they aggregate, process, and use through robustly trained algorithms. This is the backbone of platforms’ business models and with this accumulated user data they create an uneven landscape for competition with smaller tech companies in the EU and raise significant challenges in fostering an inclusive, open, and safe online environment where human rights are upheld and promoted.

The EU has demonstrated a commitment to strengthening data protection and to regulate digital platforms as a basis for safeguarding human rights and addressing the potential harms of digital platforms to societies writ large. The General Data Protection Regulation (GDPR) aimed to lay out a thorough and all-encompassing approach to data protection in the digital era, covering crucial principles such as consent, transparency, and individual rights. Its impact has even resulted in the formulation of subsequent global data privacy laws and standards. However, its enforcement has been heavily criticized because of its one-stop-shop system which allowed large platforms to get away with abusing personal data.

The EU’s Code of Conduct on Countering Illegal Hate Speech Online has established a cooperative framework among technology companies to actively address and respond to hate speech on digital platforms. It has facilitated a more organized and effective approach to recognizing, reporting, and removing hate speech. Yet it has been limited and was criticized by UN experts for undermining the Rule of Law.

The most recent attempt, the Digital Services Act (DSA), establishes new legal and regulatory norms for online platforms and intermediaries in dealing with illegal content. It also includes due diligence requirements to foster a safer online ecosystem, with the most stringent rules reserved for VLOPs. Even though the law went into effect on November 16, 2022, key elements related to risks, transparency, and stakeholder engagement have yet to be defined, posing a challenge in aligning the ongoing DSA implementation with these crucial aspects. Regardless, the DSA's implementation is in full swing.

Despite the EU's robust legislative architecture aimed at protecting users' rights online, VLOPs are still unable to fully control hateful content, misinformation and violence on their platforms, despite the policy changes and safety measures they have already introduced. **Legislative attempts to regulate VLOPs have prompted important changes, but lack guidance and standards for how to deal with systemic risks for online safety.** In the case of the DSA, the legislation has established standards and obligations for VLOPs that have changed the game on data security, user experience, and more, but they don't provide guidelines or standards for VLOPs to concretely address concepts that are integral to online safety, including conflict sensitivity.

Additionally, as has already sufficiently been documented by others, **the lack of clear definitions in the DSA regarding crucial components such as transparency, stakeholder engagement, and risk, has serious real-world implications.** This lack of detail results in confusion among tech companies on how to handle these aspects, leading to inconsistent application and potentially ineffective compliance. This ambiguity may also result in differing interpretation and implementation across EU member states, resulting in a fragmented regulatory environment. It also makes it hard to accurately assess the DSA's impact in addressing hate speech, misinformation, and online violence. Without well-defined parameters, measuring the DSA's effectiveness becomes challenging, making it difficult to enhance its impact in these critical areas.

What has not been looked into yet, is the impact that the DSA and its attempt to regulate tech companies behavior could have on online/offline violence in the rest of the world, including in conflict-affected countries, and conflict-affected and fragile settings, nor which lessons can be learnt from the rest of the world to improve the implementation of the DSA in Europe and beyond. This report seeks to address these gaps.

The report first looks at why the European Commission should care about conflict-sensitivity and what it is, before it examines how conflict sensitivity has been handled by three very large social media platforms, Facebook, Snapchat, and TikTok, based on publicly available info on their first risk assessments. It then shares lessons learned from initiatives in Sri Lanka, Kenya, and Democratic Republic of Congo to create a healthy online environment for users of VLOPs, and ends with reflections on what the EU could do next to better integrate these lessons in the implementation of the DSA.

Why should the European Commission care about conflict-sensitivity?

Plenty of evidence points to how VLOPs can become a breeding ground for hate speech, extremist ideologies, and violent content and these challenges create divides based on religion, race, climate change, generational differences, social classes, and immigration status, both on an international and local level. This creates a toxic digital environment that poses a significant threat to public safety and social cohesion, amplifies radicalization, and fuels real-world violence and polarization by providing a virtual echo chamber for polarizing and extremist views.

First, drawing on media reports from the first risk assessments, most Facebook, Snapchat, and TikTok users in the EU are teenagers between the ages of 13-17, a critical category that the DSA is keen to protect online. According to a 2023 EU Parliament report, this group is most vulnerable to experiencing the damaging effects of online violence which may negatively affect their mental health and general wellbeing. They are also the easy targets for political or ideological manipulations, which can impact democratic processes and public discourse. By their sheer numbers online, the dissemination of misleading information and disinformation by youth can quickly erode trust in online platforms, jeopardizing their credibility and resulting in a loss of user confidence and engagement.

Second, despite ongoing efforts to improve content moderation on these platforms, they are replete with mis/disinformation and hateful content during elections and times of major political crises. An EU report revealed that VLOPs, including Meta and TikTok, failed to constrain a massive Kremlin disinformation campaign during the first year of Russia's invasion of Ukraine. The study found that "the reach and influence of Kremlin-backed accounts has grown further in the first half of 2023, driven in particular by the dismantling of X's safety standards."

Third, with several countries in the EU heading for elections in 2024, the risk posed by foreign interference in election outcomes is significant and requires urgent actions to address the current vulnerabilities of social media platforms to this risk. Similarly with countries such as Bulgaria, Denmark, Finland, France, Germany, Greece still struggling with far right-wing groups, the risk for online manifestations of conflicts to escalate in violence offline is imminent. These groups exploit anti-immigration sentiments and xenophobic stereotypes in social media platforms to mobilize targeted attacks against specific groups of foreigners and ethnic minority groups who are EU nationals. In Greece, for example, online and offline violence towards 2nd generation Afro-Greeks by Golden Dawn is a major problem. This is an ever-present systemic risk that the DSA ought to require platforms to identify and proactively develop mitigation measures. Just like any other form of extremism, right-wing extremism can spread hateful language, discrimination, and encourage violence. This can cause disruption to social harmony and put targeted communities in danger. The DSA can promote a safer online environment by both concretely tackling illegal hate speech and incitement to violence, whilst also forcing platforms to better address harmful content that also puts these groups at risk.

As the DSA becomes a powerful driving force on tech regulation inside the EU and is expected to influence corporate and government behavior in the tech sphere globally, it is essential for the EU Commission and DG-CNECT to also pay attention to international trends and lessons–including from conflict-affected countries– that can contribute to strengthening the DSA's implementation. This requires a commitment to formally seek out information from these contexts and an understanding of ways they can contribute to a healthy online ecosystem for EU users.

When operating in conflict-affected areas, VLOPs and other content-sharing websites confront major hurdles in managing violence, misinformation, and hateful content. The huge scope and reach of platforms, as well as the lack of effective conflict-sensitivity methods to address potential unintended outcomes such as amplification of violence, add to the challenges of monitoring harmful content in contexts of ongoing violence.

Conflict-affected contexts provide fertile ground for the dissemination of rumors, fake news, hate speech, and misinformation campaigns, all of which can exacerbate conflicts, impede peace efforts, and further endanger vulnerable populations. Online platforms often serve as virtual battlegrounds for different factions in conflicts, and due to the sheer scale and virality of content, they frequently struggle to handle mis/disinformation efficiently. Platforms must strike the correct balance between protecting the right to freedom of expression and access to information, whilst increasing capacity for responsible content moderation and due diligence in these situations. In some circumstances, insufficient content moderation may allow illegal content and harmful speech to spread rapidly, dangerous content to grow, while overly strong moderation may be regarded as biased censorship, fuelling tensions even more.

Efforts to address these challenges should include a commitment to conflict-sensitive platform design and content moderation. Platforms can invest in training content moderators to understand the nuances of conflict-affected areas, engage with local experts and civil society organizations, and incorporate context-specific insights into their policies and practices. Furthermore, **increased transparency and collaboration with relevant stakeholders, including governments, NGOs, and international organizations, can facilitate a more comprehensive and informed approach to mitigating violence, misinformation, and hateful content in conflict-affected contexts**.

Drawing on the publicly available media reports on the first risk assessment submitted by the three large platforms to the European Commission, **none of the platforms spelled out conflict sensitivity as an important criteria to identify and mitigate risks in the design and content moderation of their platforms**. Against this backdrop, a conflict sensitivity requirement in the DSA's risk assessment could improve both the human rights protection of EU users, particularly teenagers, and help establish mitigation measures against potential violence linked to online polarization ahead of major elections and political crises.

What is conflict-sensitivity for tech?

Some of these impacts could be avoided by building on best practice from other places around the globe, and by making conflict sensitivity a requirement for VLOPS. There are several ways for defining conflict sensitivity and it could sometimes be perceived as a catch-all phrase. At Search for Common Ground, we define conflict sensitivity as the understanding that all interventions interact with conflict dynamics and should seek to avoid aggravating conflict while maximizing their positive impact. Business for Social Responsibility defines conflict sensitivity for tech platforms as the ability of online platforms to understand, adapt to, and mitigate the risks and impacts of their operations on conflicts and human rights in areas experiencing conflict or political instability. It encompasses efforts to minimize harm, promote peace, and prevent the exacerbation of conflicts through content moderation, community guidelines, and algorithmic decision-making.

Unlike general human rights and risk assessment frameworks, conflict sensitivity for tech platforms necessitates a deep understanding of the local context, including the root causes, dynamics, and actors involved in conflicts. It requires platforms to identify how their operations can influence or be influenced by ongoing conflicts.

Concretely, in the DSA's implementation context, conflict sensitivity highlights the crucial importance of requiring VLOPs to understand how their platforms' design and content moderation practices address or exacerbate online and offline manifestations of social or political tensions in specific EU member States. **The conflict sensitivity approach consists of 3 key steps:**

Understanding the contexts in which tech platforms operate requires understanding existing peace and conflict dynamics and the interests and incentives of key actors.

Assessing how platforms' design and content moderation policies might impact social cohesion, conflict, unpacking risks and opportunities.

Adapting policy interventions to minimize harm, maximize opportunities to build social cohesion and stability, and adapt to evolving conflict dynamics.

Search for Common Ground considers the following three elements as core to an effective approach to conflict sensitivity in online platforms: trust, agency, and horizontal cohesion.

- Trust: Trust is the cornerstone of any effective conflict sensitivity framework. Ideally platforms must foster trust among users, governments, and civil society.

- Agency: Empowering users and stakeholders with increased agency in shaping digital environments is vital for conflict sensitivity.

- Horizontal Cohesion: Horizontal cohesion refers to the collaboration and coordination among stakeholders in addressing conflicts.

These elements can be measured very clearly in alignment with the Peace Impact Framework which supports our understanding of impact in conflict and peace dynamics both on and offline.

If we apply these to ensure a conflict sensitive approach to risk assessment by VLOPs, then the following five questions should be asked:

- How do platforms identify drivers of online and, subsequently, offline manifestations of conflicts?

- How do they contribute to addressing them?

- How does platform design and content moderation advance trust, agency, and horizontal cohesion in times of political unrest and conflict?

- Who are the main vulnerable groups? How are they identified? Is the selection process inclusive?

- What unintended consequences could the platforms' design and content moderation practices have on conflict dynamics? What mitigation measures will be put in place to address these harms proactively?

A conflict sensitivity lens in online platforms goes beyond 'doing no harm', and explores how Tech regulation and design can do more good for societies, considering that each context presents a unique challenge and no one-size-fits-all risk assessment approaches will succeed in the long haul. This requires a regular conflict analysis to establish a clear understanding of local dynamics. Furthermore, the capacity to understand the ever-shifting terrain of risk in conflict-affected contexts is critical for supporting and upholding human rights online.

Do’s and Don’ts for VLOPs in a Risk Assessment

Based on learning from Human Rights Impact Assessments by platforms prior to DSA enforcement, engagement in Meta’s Trusted Partnership Program, insights from ongoing debates on risk assessments in Brussels, and policy reports on the same from various actors including Penn America and Meedan, Tech Policy Press, AccessNow and ECNL, we identified five key principles that can serve as guidelines for VLOPs, ensuring that their presence in conflict zones does not accidentally stoke the flames of discontent or further inequity:

1. Identifying and mitigating risks:	
• What works: To identify harmful content, social media companies use a combination of *proactive* detection via automation and human moderation, and *reactive* detection via user reporting, which automated systems or human moderators then adjudicate. Human moderators are better equipped to consider the nuances of language and cultural and sociopolitical context.	• What doesn’t: Due to platforms’ internal considerations and decision-making processes, there are often delays in removing certain flagged harmful content –which violate a platform’s terms of service and may continue causing harm to societies. Furthermore, content moderation exerts a heavy emotional toll on human moderators who are often economically exploited and generally from conflict-affected countries. Platforms should not put in place mechanisms for human content moderation without providing the necessary holistic support for these teams.

<table>
<tr><th colspan="2">2. Collaboration and information sharing:</th></tr>
<tr><td>• What works: Collaborative efforts such as the Global Internet Forum to Counter Terrorism (GIFTC), Global Network Initiative (GNI), the Council on Tech and Social Cohesion (CTSC) among others are helpful multistakeholder efforts enabling alignment on reporting mechanisms, platform policies, designing tech for cohesion, and the development of shared best practices among tech platforms, regulatory bodies, governments, and civil society actors.</td><td>• What doesn't: Limited collaboration and information silos hinder a holistic understanding of risks and may lead to inefficiencies in the risk assessment. For example, the Code of Practice but also the Code of Conduct on illegal hate speech have both been heavily criticized for the information silos, lack of concrete engagement with civil society and the result being a somewhat weak Code of Practice and a Code of Conduct that has been criticized by the UN for undermining rule of law and potentially encouraging censorship.</td></tr>
<tr><th colspan="2">3. User Involvement and Feedback:</th></tr>
<tr><td>• What works: The DSA's provisions 14 to 24 advance greater user control (opt-out by default, control of recommender systems functions, dispute settlement bodies etc) as they report abuse. Actively involving users in the risk assessment, seeking their feedback, and considering their experiences can provide unique perspectives on platform usage and potential risks. Facebook also claims to have increased user empowerment through their recently launched 22 system cards, which helps users to understand how AI systems rank content for feeds.</td><td>• What doesn't: As pointed out in several studies, including this June 2020 study commissioned by the European Parliament, mechanisms used to flag harmful content on various social media platforms have limited user engagement and feedback loops. Many users don't understand the reporting process, including where they stand, what to expect after submitting a report, and who will review it. Furthermore, users are often left in the dark about whether a decision has been made regarding their report and why. Important risks and concerns may be overlooked if user feedback is ignored or user insights are not considered.</td></tr>
</table>

4. Agility and Adaptability	
• What works: Offering users more comprehensive reporting options. For example, an expedited reporting process is linked with Trusted Flaggers in the context of the DSA. When reporting harmful content, sometimes users have different—and occasionally competing—needs. In some situations, users want the reporting process to be quicker and easier, especially if they are experiencing a high volume of abuse. In other situations, users want a more comprehensive reporting process that provides room for context, especially if they are being harassed across multiple platforms or combined with offline threats or abuse or if the abuse is coded or otherwise requires additional information.	• What doesn't: Rigid risk assessment frameworks that cannot swiftly adapt may become outdated and less effective in mitigating evolving risks.
5. Transparency and Accountability:	
• What works: Making transparent to users platforms' rationale on decision-making on whether to keep or remove flagged content. For example, the recently created transparency dashboard by the EU Commission is a good baseline model for this. It mandates that online platforms operating in the EU provide access to all content moderation decisions thus empowering users to track reports, outcomes, and history of flagged content on very large online platforms.	• What doesn't: Secrecy and lack of transparency and accountability on platforms' decision-making processes on flagged content can lead to skepticism and distrust, potentially undermining risk assessment efforts. Many platforms have inboxes for reports, but they can be hard to find, and communication is limited. Instagram, Facebook, and TikTok have inboxes that are not easily accessible. Twitter sends updates via notifications and email. YouTube has a "Report history" section, but it only tracks reported videos, not comments. It also doesn't show the report's progress or all the submitted information.

Reporting mechanisms on social media platforms need significant improvements to better protect users and uphold free expression online. We need more accessible, user-friendly, efficient, effective, and transparent reporting features. However, progress has been fragile and insufficient, with some platforms hiring fewer employees for Trust and Safety teams. Clear standards on minimum viable reporting systems are needed to protect users and free expression.

Case studies: Kenya, Sri Lanka, and the Democratic Republic of Congo

Beyond what we identified as do's and don'ts in the risk assessments, we should also look beyond Europe for ways in which the European Commission can enforce effective risk assessments and action of VLOPs under the DSA, with a focus on conflict-sensitivity, safeguarding minors' online rights and fostering trust, agency, and horizontal cohesion. From our work as Search for Common Ground, we see some valuable lessons from Kenya, DRC and Sri Lanka.

Multi-stakeholder Approach in Assessing Online Risks: Kenya elections 2022

Social media platforms emerged as fertile grounds to amplify politically instigated hate speech, disinformation, misinformation, and manipulation around the August 2022 elections in Kenya. In partnership with Build Up, using social media listening, Search for Common Ground and a group of local CSOs monitored online conversations, analyzed 4133 Tiktok videos and over 140000 Facebook posts in two counties to track hate speech, disinformation, and misinformation trends to understand how these could potentially affect conflicts offline. Crowdtangle was used to gather data from Facebook, while TikTok data was scraped from Google Chrome through HTTP Archive format files. We found out that (1) there were strong signs of electoral divisions along ethnic and racial lines, (2) inflammatory content targeting political competitors was prevalent on both platforms, and (3) women contending for political positions were more targeted by hate speech and disinformation campaigns.

As a result, and parallel to the social media listening exercise, Search for Common Ground established an Early Warning and Early Response system between communities, civil society organizations and government authorities to identify and address imminent threats to peaceful elections, particularly looking at conflict trends, drivers and locations that could potentially trigger or experience violence. Social media listening and the EWER system were identified catalysts to offline violence experienced during the campaign period and the election day in Mombasa, precisely in Nyali and Mvita. Where supporters of opponents were engaged in physical fights throughout the campaign and on election day. To mitigate these problems, Search for Common Ground worked in partnership with Facebook to identify problematic content and also reached out to TikTok for the removal of channels that were guilty of publishing and promoting misinformation. While TikTok's response was good, not all the problematic content violating the platform's community standards was removed from the platform. This was due to the nature of the language used and self regulation of content. To contribute in addressing this challenge, Search for Common Ground decided to train social media users to de-escalate the spread of hate speech and provide alternative narratives.

The Kenya case study illustrates the real-world challenges associated with the spread of hate speech, disinformation, and misinformation on social media platforms during a politically charged event, in this case, the August 2022 elections in Kenya. It showcases the impact of such online content on offline conflicts and electoral violence. In connection with the DSA, the case surfaces the following learnings:

- **Proactive and inclusive risk assessments:** The Kenya case study highlights the value of proactively engaging multiple actors (including CSOs) in conducting risk assessment to prevent the spread of dis/misinformation and hate speech online in times of elections. As required by Article 34 of the DSA, VLOPs should conduct risk assessments to address systemic risks to fundamental rights on their platforms. This is particularly relevant in the context of major elections taking place in 2024. The September 2023 elections in Slovakia revealed significant platform vulnerabilities (Alphabet, Meta, and Tik Tok) for hate speech, disinformation, and pro-Russia propaganda. Against this backdrop, it is crucial that future risk assessments by VLOPs are conducted with a multi-stakeholder framework to mitigate the impact of risks on EU's democracy.

- **Real-world application**: In the Kenyan example, social media platforms, particularly Facebook and TikTok, became channels for the amplification of hate speech and disinformation during the election period. This content had the potential to incite offline violence, as evidenced by physical fights between supporters of different political opponents. Here in Europe, in 2021 one in every five Roma, Europe's largest ethnic community, reported receiving hate-motivated harassment both online and offline. As also underscored by the recent report of AccessNow & ECNL on risk assessment, anti-gypsyism's broad silencing effect on the internet creates barriers to Roma people's involvement in public life and the use of the internet and social media. Similarly, future risk assessments should be guided by guidelines informed by a deeper understanding of the specific online and offline experiences of marginalized groups to protect them from online and offline violence. The Kenya case study also demonstrates how politically instigated hate speech, disinformation, and misinformation and other more pernicious content like fear speech present real-world risks to trust, agency, and horizontal cohesion, which aligns with the DSA's concern about identifying and mitigating risks.

- **Parallel efforts**: In response to these issues, the Kenya case study describes parallel efforts to address systemic risks on platforms. This includes social media listening, an Early Warning and Early Response (EWER) system, and direct engagement with the social media platforms. Considering platforms' limitations with automated and human content moderation, a conflict-sensitive approach to risk assessment requires implementing additional strategies, including awareness campaigns, to mitigate the spread of online harms. These parallel actions support the obligations placed on very large online platforms by the DSA.

- **Training and de-escalation:** The Kenya case study also highlights the need to advance user empowerment by training social media users to de-escalate the spread of hate speech and mis/disinformation. This aligns with the broader approach of the DSA, which encourages platforms to implement measures to address harmful content proactively.

More user empowerment for minors, youth, and other marginalized groups: Sri Lanka 2019.

On April 21, 2019, Sri Lanka experienced a series of coordinated suicide bombings on churches and luxury hotels nationwide. The attacks were carried out by a local Islamist extremist group, later identified as the National Thowheed Jama'ath (NTJ), resulting in the tragic loss of over 250 lives and hundreds more injured. The targets included churches in Colombo, Negombo, and Batticaloa during Easter Sunday services and several high-end hotels in the capital. Following the Easter bombings, a false rumor that 11 police officers had been killed in a Muslim town went viral on social media resulting in a government minister appearing on television and repeating the rumor. Understanding the danger and potential damage that this rumor would have on trust, agency, and horizontal cohesion among Sri Lankans in an already fragile environment, Search for Common Ground's trained community stewards – youth who review user generated content on social media– mobilized immediately. These community stewards mobilized all of the groups that they moderated online and uploaded fake news banners. This was so effective, that 4 hours later, the minister returned to television to acknowledge his mistake for spreading misinformation.

Furthermore, to promote pluralism and peace through social media, Search for Common Ground in Sri Lanka has formulated a concept called 'cyber guardians' which empowers youth to combat hate and fake content in cyberspace. The project mainly targets four districts: Colombo, Puttalam, Kandy, and Batticaloa, which were identified as districts with widespread racially and religiously motivated hate speech on social media.

In terms of identifying and mitigating risks of online polarization on platforms, CiTW Search for Common Ground is leading a two-step process to engage with Big Tech. The first step involves consulting with CSOs in Sri Lanka to develop a tailored strategy for dialogue with each Big Tech company separately. The second step involves engaging with each Big Tech using an evidence-based approach based on the developed strategy and including clear asks. As part of this project, Search for Common Ground's partner in Sri Lanka, Hashtag generation, monitors social media platforms for hate speech and flags harmful content to platforms. While platforms' content moderation policies differ, and platforms' decision-making criteria on the removal of content remains nebulous for the most part, this tailored strategy for dialogue with each Big Tech company, which is also embedded in the conflict sensitivity approach, is enabling constructive advances in addressing online hate in Sri Lanka. This effort is part of a broader push for Big Tech companies to sign a Code of Practice. At the same time, it is important to note that CSOs are against national legislation, due to past experiences of using laws to curtail freedom of speech and expression.

- **Conflict Sensitivity Approach:** The case study mentions that the strategy for engaging with Big Tech companies is embedded in a "conflict sensitivity approach." This approach aligns with the DSA's goal of addressing online risks that can contribute to real-world conflicts and harm social cohesion.

- **Avoidance of National Legislation**: The case study notes that CSOs in Sri Lanka are against national legislation to address online hate due to concerns about freedom of speech. This aligns with a key aspect of the DSA, which seeks to establish a comprehensive European framework to avoid a patchwork of national laws, ensuring consistent standards for online content while respecting fundamental rights.

- **VLOPs and Responsibility**: The case study highlights the spread of false rumors and misinformation on social media platforms, which can have real-world consequences, including harm to horizontal cohesion and trust. It also underscores strategies by local actors to advance platform responsibility to address online hate and misinformation. Search for Common Ground's efforts in Sri Lanka to monitor social media for hate speech and harmful content, can be seen as a proactive approach to identifying and mitigating such risks, which is an important goal in the DSA implementation. Furthermore, Search for Common Ground's two-step process involving consultations with civil society organizations (CSOs) to develop tailored strategies for dialogue with each Big Tech company is an example of how to organize meaningful consultations between CSOs and Tech actors. This example underscores also the importance for the EU Commission to engage in a policy dialogue with expert CSOs in conflict-affected countries with the aim to draw lessons and best practices on addressing online harms.

Building on CSO's expertise in moderating content: DRC elections 2023

The Democratic Republic of Congo (DRC) has a history of deeply rooted tensions and online conflicts during elections, which could pose significant risks for the upcoming December 2023 elections. Some of these risks include hate speech between political opponents, the spread of polarizing misinformation, tense relations with neighboring Rwanda, possible internet shutdowns, and censorship of freedom of expression. These factors contribute to a polarized environment that could make it difficult to hold peaceful elections.

Search for Common Ground conducted consultations with various stakeholders, including Facebook representatives, journalists, bloggers, civil society organizations (CSOs), and members of the Congolese diaspora in Brussels to assess the risks of online polarization and conflicts during the upcoming December 2023 elections. With other CSOs who are part of Meta's Trusted Partnership Program (TPP), Search for Common Ground is asking Facebook to support the monitoring of online trends and risks of online polarization in the weeks ahead of the general elections and provide technical and financial support to certified fact-checkers in the DRC to develop a central website with verified information on the upcoming DRC elections.

Lastly, Search for Common Ground also recommended the inclusion of trained journalists and bloggers in the EU's electoral observation team in the DRC to provide expert insight into the effects of hate speech and disinformation campaigns during elections. These efforts demonstrate the value of working closely with expert CSOs, bloggers, and journalists in conflict-affected countries in addressing mis/disinformation and could serve as a model for how platforms in Europe can in practice identify potential manifestations of conflicts on their platforms and mitigate their societal impacts. For the CSO partnership to work effectively, there must be clear transparency on how and when their input is taken into account by platforms.

In connection with the DSA, the DR of Congo case study highlights the following lessons and best practices:

- **Wider consultations with stakeholders**: The value of meaningful public consultations in future risk assessments cannot be underscored enough. Since October 2022, Search for Common Ground, in collaboration with various stakeholders, has conducted consultations to assess the risks of online polarization and conflicts during the elections. This regular multi-stakeholder exercise enables an environment of trust-building among actors and helps deconstruct the us vs them false binary that is still prevalent in the DSA's implementation discourse.

- **Expert insight:** In the case study, Search for Common Ground, along with other CSOs in Meta's Trusted Partnership Program (TPP), is requesting support from Facebook to monitor online trends and risks of online polarization, as well as to provide support to certified fact-checkers in the DR of Congo ahead of elections. While TTP is fraught with its challenges as highlighted by Internews' August 2023 report, it is a best practice for platforms to improve mechanisms for receiving and applying expert insight on their product design and content moderation policies. Furthermore, the case study also recommends the inclusion of trained journalists and bloggers in the EU's electoral observation team to provide expertise on the effects of hate speech and disinformation campaigns during elections. Expert groups, as mandated by Article 3 of the Commission Decision can also assist with risk assessment and engagement processes, and as recommended by ARTICLE 19's 2023 report on collaborations between the European Commission and civil society. Ahead of the 2024 EU elections, ensuring meaningful participation of trusted expert CSOs, journalists, bloggers, and research institutions can enhance the DSA's efforts to address harmful content.

- **Transparency and Accountability:** The case study emphasizes the need for clear transparency on how and when input from CSOs and other stakeholders is considered by platforms. Equally, in platforms' product design, the European Commission, and in particular DG-CNECT should require platforms to publicly provide product experimentation results on outcomes of societal interest for any meaningful product design decision. Since products and technology are continually evolving, we need this level of transparency to meaningfully understand the causal impact of future product decisions (e.g., optimizing for time spent or comments consumed) on outcomes of interest in conflict (e.g., predicted hate speech, views of content reported for violence incitement). Furthermore, providing product experimentation results would help disentangle what platforms are hosting (they didn't create this conflict), from what their product choices are causing (as measured experimentally, which is the scientific community's method for adjudicating causality).

The DSA and non-EU actors

As we have seen from media reports on the first tranche of “risk assessments’ turned in by VLOPs in August 2023, this requirement has resulted in product changes that are helpful. For example, Facebook has recently introduced an ‘Ad Library’ in an effort to increase transparency of ads targeting EU users, along with dates the ad ran, the parameters used for targeting (e.g. age, gender, location) etc. Still, tech products are deployed with an *inadequate understanding and mitigation of region-specific risks, limited effectiveness in curbing hate speech and misinformation, and a failure to address marginalized communities’ unique vulnerabilities*.

To refine and improve the DSA, the EU Commission can learn from successful international implementations of transparency, stakeholder engagement, risk assessment, and related measures. Adapting and tailoring the DSA to align with proven strategies from other countries increases its effectiveness and efficiency. Studying these experiences also helps identify potential pitfalls and areas for improvement, ensuring that the DSA remains dynamic and up-to-date with evolving tech landscapes. **Additionally, incorporating these experiences fosters international cooperation and harmonization**, enabling a more consistent and impactful approach to addressing online challenges across borders. Ultimately, integrating international lessons strengthens the DSA’s potential to create a safer, more transparent, and user-friendly online environment for individuals within the EU and beyond.

Similarly, the EU Commission should pay more attention to the DSA’s impact on the rest of the world, mainly in conflict-affected countries, which often carries the heaviest brunt of the negative effects of tech design and regulation in their societies, where companies do not invest as much resources for content moderation, and where several countries still lack needed infrastructures in place to protect fundamental human rights online. Toward this end, **the EU Commission and DG-CNECT need to engage and sustain a policy dialogue with conflict-affected countries on the DSA and its subsequent legislations to advance VLOPs’ compliance to the DSA’s obligations**.This call for the EU’s policy dialogue with conflict-affected countries is supported by both the EU’s Digital Diplomacy as well as an analysis of the extraterritorial implications of the DSA. **Non-EU actors can be directly involved in the DSA in seven ways:**

- **Representation:** non-EU stakeholders can represent service recipients established or located in the EU to lodge a complaint to a digital services coordinator. Concretely, an organization based in Syria would be able to file a complaint on behalf of Syrian citizens in the EU, but an organization based in Berlin that represents the Rohingya in Bangladesh would not be able to file a complaint on their behalf.
- **Vetted researchers:** non-EU stakeholders could apply to become vetted researchers
- **Auditors:** non-EU stakeholders could be appointed as an auditor by a VLOP
- **Risk assessment process**: non-EU stakeholders could be part of the risk assessment process at the invitation of a VLOP
- **Codes of conduct:** non-EU stakeholders could be involved in drawing up codes of conduct
- **Crisis protocols:** non-EU stakeholders could be involved in drawing up
- **Independent expert or auditor:** non-EU stakeholders could be required to provide information to DSCswhich would then allow them to submit written comments to the DSC or the Commission in relation to the case at hand.

The most promising way for non-EU actors to be involved in the DSA is indirectly, through providing evidence and cases related to the VLOPs risk assessment provisions. All this is to emphasize the necessity for the EU to meaningfully and actively seek out conflict-affected countries' perspectives in the continued implementation of the DSA and future related legislations.

Conclusion

Based on lessons learned from the first round of risk assessments of VLOPs as well as Search for Common Ground's experience in addressing mis/disinformation and hate speech in Kenya, Sri Lanka, and the DRC, we can offer the following five recommendations to the European Commission, the DSA enforcement team, and DG-CNECT:

- The newly established DSA enforcement team of the European Commission should include conflict sensitivity as a core requirement into risk assessment guidelines
- The newly established DSA enforcement team of the European Commission should enable a multi-stakeholder process (i.e with civil society, industry, and EU input) to formulate risk-assessment guidelines and ensure a meaningful participation of conflict-affected communities in developing these guidelines.
- The European Commission, and in particular DG-CNECT, should engage and sustain a policy dialogue with CSOs in conflict-affected countries on the DSA by conducting quarterly public consultations with relevant authorities and CSOs in these countries on addressing systemic risks on VLOPs
- The European Commission, and in particular DG-CNECT, should require platforms to publicly provide product experimentation results on outcomes of societal interest for any meaningful product design decision.
- The European Commission should ensure that the impacts of recommender systems, as a crucial design layer of large social media platforms, are assessed comprehensively from a conflict-sensitive and human-rights perspective.

Conflict sensitivity is vital to the Digital Services Act's mission to create a safer online environment. While the first risk assessments submitted by very large online platforms represent a step in the right direction, there is room for improvement. Striking a balance between automated moderation and human judgment, enhancing transparency and accountability, and embracing cultural sensitivity is crucial. Moreover, active engagement with civil society and rapid responses to emerging conflicts can further strengthen these assessments. Online platforms can play a pivotal role in promoting a safer and more harmonious digital space by continually refining their conflict-sensitive approaches.

Sources

1. Final text of the DSA https://www.eu-digital-services-act.com/Digital_Services_Act_Preamble_31_to_40.html

2. The GDPR https://gdpr-info.eu/

3. The Code of Conduct on Countering Illegal Hate Speech Online https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

4. David Sulivan and Jason Pielemeier, 'Unpacking 'Systemic Risk' under the EU's Digital Services Act, July 2023 https://techpolicy.press/unpacking-systemic-risk-under-the-eus-digital-service-act/

5. The Influence of Social Media on Development of Children https://www.europarl.europa.eu/RegData/etudes/STUD/2023/733109/IPOL_STU(2023)733109_EN.pdf

6. Application of the Risk Management Framework to Russian Disinformation Campaigns https://op.europa.eu/en/publication-detail/-/publication/c1d645d0-42f5-11ee-a8b8-01aa75ed71a1/language-de

7. Mirza Buljubašić, 'Violent Right-Wing Extremism in The Western Balkans', 2022 https://home-affairs.ec.europa.eu/system/files/2022-08/ran_vrwe_in_western_balkans_overview_072022_en.pdf

8. Fahrinisa Campana, 'The Don't Accept You: 'Afro-Greeks Struggle to be Seen', November 2020 https://www.aljazeera.com/features/2020/11/18/afro-greeks

9. Ayesha Khan, 'Conflict-sensitive Human Rights Due Diligence for Tech Companies', December 2022 https://www.bsr.org/en/blog/conflict-sensitive-human-rights-due-diligence-for-tech-companies

10. Pen America, 'Shouting into the Void', June 2023 https://pen.org/report/shouting-into-the-void/

11. ECNL & AccessNow, 'How Tech Corporations like Google, Meta, and Amazon should Assess Impacts on Our Rights.', October 2023 https://edri.org/our-work/how-tech-corporations-like-google-meta-and-amazon-should-assess-impacts-on-our-rights/

12. Alexandre Destreel, 'Online Platforms' Moderation of Illegal Content Online', June 2020https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf

13. Punjajoy &al. 'On the Rise of Fear Speech In Online Social Media', November 2022 https://www.pnas.org/doi/epdf/10.1073/pnas.2212270120

14. Dr. Suzanne Vergnolle, 'Putting Collective Intelligence to the Enforcement of the Digital Ser-

vices Act,' May 2023 https://dsa-enforcement.vergnolle.org/

15. Internews, 'How to Save Meta's Trusted Partner Program', August 2023 https://internews.org/resource/safety-at-stake-how-to-save-metas-trusted-partner-program/

16. Nathaniel Lubin and Ravi Iyer, 'How Tech Regulation Can Leverage Product Experimentation', July 2023 https://www.lawfaremedia.org/article/how-tech-regulation-can-leverage-product-experimentation-results

17. JJiaxuan Yue, Habibou Bako, Kelsey Hampton, Katie Smith, 'Online conflicts in the Sahel report' July 2022 https://cnxus.org/wp-content/uploads/2022/07/Issue-Brief_-Conflict-and-the-Online-Space-in-the-Sahel-July-2022-1-1.pdf

18. Laureline Lemoine & Mathias Vermeulen, 'The Extraterritorial Implications of the Digital Services Act', November 2023 https://dsa-observatory.eu/2023/11/01/the-extraterritorial-implications-of-the-digital-services-act/

19. Institutional Learning Team, 'Community Stewards and Social Cohesion in Digital Spaces.' Search for Common Ground, January 2023. https://www.sfcg.org/wp-content/uploads/2023/02/SFCG-Community-Stewards-and-Social-Cohesion-in-Digital-Spaces.pdf

20. Social media listening report https://cnxus.org/resource/social-media-listening-analysis-uchaguzi-bila-balaa-kenya-october-2022/

21. 4Meta's media report on the August 2023 quarterly risk assessment: https://about.fb.com/news/2023/08/new-features-and-additional-transparency-measures-as-the-digital-services-act-comes-into-effect/

22. SnapChat's media report on the August 2023 quarterly risk assessment:https://newsroom.snap.com/en-GB/digital-services-act-snap

23. Google's media report on the August 2023 quarterly risk assessment: https://blog.google/around-the-globe/google-europe/complying-with-the-digital-services-act/

24. Microsoft's media report on the August 2023 quarterly risk assessment:

25. https://blogs.microsoft.com/eupolicy/2023/08/25/microsoft-digital-services-act-online-safety/

26. Whatsapp's media report on the August 2023 quarterly risk assessment:

27. https://faq.whatsapp.com/781249240131848

28. Gabby Miller, 'Who's Afraid of the DSA', August 2023 https://techpolicy.press/whos-afraid-of-the-dsa/

29. Priyanka Shankar, 'What Impact Will the EU's Digital Services Act Have', August 2023 https://amp-dw-com.cdn.ampproject.org/c/s/amp.dw.com/en/what-impact-will-the-eus-digital-services-act-have/a-66631337

30. Addicting algorithms

31. https://www.linkedin.com/pulse/addicting-algorithms-how-digital-services-act-affects-carlo-prato?utm_source=share&utm_medium=member_ios&utm_campaign=share_via

32. Allessandra D'Angelo, 'Le Regime UE Orwellien de Censure de l'Internet', Aout 2023

33. https://pan.be/article/digital-services-act-le-regime-ue-orwellien-de-censure-de-linternet-764

34. Peace Impact framework https://cnxus.org/peace-impact-framework/

35. Iverna McGowan, Asha Allen, 'Fostering Responsible Business Conduct in the Tech Sector: The Need for Aligning Risk Assessment Transparency and Stakeholder Engagement Provisions Under the EU Digital Services Act' August 2023 https://cdt.org/insights/fostering-responsible-business-conduct-in-the-tech-sector-the-need-for-aligning-risk-assessment-transparency-and-stakeholder-engagement-provisions-under-the-eu-digital-services-act-with-the/

36. Moderating Online Content: Fighting Harm or Silencing dissent, July 2021 https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent

37. Bamako Forum https://www.sfcg.org/bamako-forum/

38. Trusted Partnership Program https://transparency.fb.com/policies/improving/bringing-local-context

39. Council on Tech and Social Cohesion https://techandsocialcohesion.org/

Search for
Common Ground
Trust, Collaboration, Breakthroughs

twitter.com/SFCG_
youtube.com/sfcg_
facebook.com/sfcg.org
instagram.com/sfcg_
searchdigitalpeacebuilding.org