CEP POLICY PAPER

The EU Digital Services Act (DSA)
Recommendations For An Effective Regulation Against Terrorist Content Online

September 2020

Alexander Ritzmann, Prof. Dr. Hany Farid, Dr. Hans-Jakob Schindler

COUNTER EXTREMISM PROJECT

Key lessons learned and recommendations for the DSA

Stop algorithmic amplification of illegal content

Algorithms on social media and video sharing platforms still do prioritize harmful or illegal content, and in some cases are even amplifying terrorist and violent extremist propaganda. Company transparency reports therefore need to include detailed information about if later on blocked or removed terrorist content had been promoted by the platform's algorithms before, including the number of views as well as data on how often the content was recommended to users.

Mandate understandable transparency

Content moderation is currently a "black box". The public is asked to trust the companies and their so called "transparency reports", which are a non-verifiable form of self-reporting. A more transparent reporting and compliance system must include explanations of the individual automated moderation tools, the technical compliance system as a whole, a better understanding of the application of moderation policies in practice and independent as well as capable oversight.

Learn from the NetzDG: "notice and take down" is insufficient

When tested by CEP, "notice and take down" systems seemed not to work properly and even when they did, failed to achieve the objectives set by the current DSA proposals. The idea that users first need to be exposed to harmful content and then have to notify for-profit companies to act, is not only hard to console with duty of EU policy makers to protect citizens from preventable harms. It also ignores the fact that bigger companies are very capable to "police" their platforms based on commercially motivated decisions concerning unwanted but legal content.

Require proactive search for illegal content

The protection of EU citizens from illegal content cannot be left to the companies, users, a few police as well as internet referral units alone. Independent and capable third parties should be commissioned and financed accordingly by the companies, but also by state actors as part of their oversight responsibility. The companies use upload- and re-upload filters already. Smart regulation that focuses on transparency, verifiability and effectiveness of automated decision-making systems, would therefore protect civil liberties more than no or weak regulation.

Learn from the regulatory and compliance structures of the financial industry

The financial industry, social media, content hosting platforms as well as messenger services face similar risks as well as comparable strategic challenges when defending against the misuse of their services. Lessons learned from the global regulatory and compliance framework of the financial industry, which developed after 2001, provide valuable insights that can be used to develop a similarly effective framework for social media, content hosting platforms and messenger services.

About CEP and the authors

The Counter Extremism Project (CEP) is a not-for-profit, non-partisan, international policy organization formed to combat the growing threat from extremist ideologies. Led by a renowned group of former world leaders and diplomats it combats extremism by pressuring financial and material support networks; countering the narrative of extremists and their online recruitment; and advocating for smart laws, policies, and regulations.

Alexander Ritzmann is a Senior Advisor to CEP and to the European Commissions' Radicalisation Awareness Network (RAN) as well as an Associate Fellow at the German Council of Foreign Relations (DGAP).

Dr. Hany Farid is a global leader in digital forensics, image analysis, and human perception and serves as a Professor at the University of California, Berkeley, with a joint appointment in Electrical Engineering & Computer Science and the School of Information. Dr. Farid is also a Senior Advisor to CEP.

Dr. Hans-Jakob Schindler is Senior Director at CEP and head of its Berlin/Germany office. He is the former Coordinator of the ISIL, Al-Qaida and Taliban Monitoring Team of the United Nations Security Council.

Please direct inquiries regarding this paper or CEP's activities to Marco Macori, CEP research fellow: mmacori@counterextremism.com; phone: +49 30 300 149 3369.

Table of content

Introduction

The Counter Extremism Project (CEP) welcomes the public consultation on the Digital Services Act, the future rulebook for digital services in the European Union. As EU Executive Vice-President Margrethe Vestager states[1]: “we are committing to build a safe and innovative digital future”. The scope of the task aimed at making the internet safer for EU citizens is vast. An estimated 720,000 hours of video content are uploaded to YouTube every day, and some one billion posts, including 300 million images, are shared on Facebook each day. Untransparent algorithms structure and recommend data to the individual user with sometimes harmful consequences.

The dissemination of terrorist content is one of the most widespread and most dangerous forms of misuse of online services[2]. Since most social media and video sharing platforms are for-profit enterprises, driven and measured by their ability to increase profits and the value of the company, conflicts between different interests and objectives on how to prioritize resources and investments are inevitable. For companies to justify the allocation of significant resources towards a more effective self-policing (compliance) regime, the incentives of doing so must be of higher priority than investing those resources into the growth of the core business, which usually is selling access to user data and profiles to the highest eligible bidder.

The EU Internet Forum[3], which started in 2015, is a widely appreciated initiative by the EU Commission that has created some incentives in the form of a “if you don´t do it, we will legislate” – approach. As a result, some of the big companies allocated more resources towards protecting users from harm. Unfortunately, the produced results are not at all sufficient. The opportunity the DSA creates is to end the “all carrots, no sticks”-era for social media for-profit companies. Currently, they are still largely shielded from liability and allowed to operate in an opaque manner without any independent oversight.

In 2020, the Counter Extremism Project (CEP) Berlin carried out a sample analysis to test the extent to which YouTube, Facebook and Instagram block “manifestly illegal” terrorist content upon notification[4]. Our findings raise doubts that the currently applied compliance systems of these companies achieve the objectives of making their services safe for users. As a result, CEP proposed requirements for effective and transparent compliance regimes[5] with a focus on automated decision-making tools and lessons learned from the regulation of the financial industry.

This Policy Paper combines all the relevant lessons learned and gives concrete recommendations on how to make the DSA also an effective regulation against terrorist content online.

Lessons (to be) learned for an effective regulation against terrorist content online

1) Ending the algorithmic amplification of illegal content

Social media decide every day what is relevant by recommending content to their users. They have learned that specific content, especially of the outrageous, divisive, and conspiratorial kind, increases engagement and retention time, which allows for the maximum extraction of their user´s data[6]. Access to this user's data is then sold to the highest bidding advertiser. This is, in short, the business model of most for-profit social media companies.

The role algorithmic amplification plays in content consumption is an issue that therefore must be confronted. In March 2020, Dr. Hany Farid co-authored the study "A Longitudinal Analysis Of YouTube's Promotion Of Conspiracy Videos"[7], analyzing YouTube's policies and efforts toward curbing its algorithm's tendency to spread conspiracy theories. The researchers found that a more complete analysis of YouTube's algorithmic recommendations demonstrated that the proportion of conspiratorial recommendations is "now only 40 percent less common than when the YouTube's measures were first announced." In order to address the effect of mis- and disinformation has had on the Internet and society as a whole, all stakeholders must come together to do better. After reviewing eight million recommendations over 15 months, researchers determined that the progress YouTube claimed[8] to have achieved in June 2019 when the company stated that it had reduced the amount of time its users watched recommended videos including conspiracies by 50 percent — and the company's claim[9] in December 2019 of a reduction by 70 percent — did not make the problem of radicalization on YouTube obsolete nor fictional.

Lesson: Algorithms (still) promote terrorism

Despite public demand and some efforts by companies, algorithms still do prioritize harmful or illegal content, and in some cases are even amplifying terrorist and violent extremist propaganda. For example, according to reports more than half of users that joined a white nationalism group on Facebook had received a recommendation for this group by Facebook[10].

Recommendation:

Company transparency reports need to include detailed information concerning whether subsequently blocked or removed terrorist content had been promoted by the platform's algorithms before, including number of views as well as data on how often the content was recommended to users.

2) Learning from the German Netzwerkdurchsetzungsgesetz (NetzDG law) experience

The main content moderation feature of the NetzDG is the "notice and take down" or “notice and action procedure”, which follows the EU´s approach to intermediary liability. Article 14 of the E-Commerce Directive (ECD) largely excludes digital service providers from liability concerning illegal third-party content unless they have “actual knowledge” of the content and fail to “act expeditiously to remove or to disable access” to it[11]. “Notice and take down” is currently also the main content moderation feature in most proposals for the upcoming DSA[12].

In February 2020, the Counter Extremism Project (CEP) Berlin carried out a sample analysis[13] to test the extent to which YouTube, Facebook and Instagram block “manifestly illegal” content and characteristics of banned organizations upon notification. The results of the sample analysis indicate that "notice and take down" is not sufficient to reduce illegal content online effectively.

2.1 Lessons learned from testing “notice and take down”

Lesson: "notice and take down" systems might not work properly
YouTube only blocked or deleted 35% of the reported illegal extremist/terrorist videos, despite the fact that the notice given by CEP included information about the official government ban orders which had been also confirmed by court decisions[14]. Videos with identical content were blocked in some cases but not others, which further indicates that the applied system or process is defect.

Lesson: "notice and take down" systems, even if they work properly, might not achieve the objectives set by the current DSA proposals
Facebook blocked illegal images reported by CEP but did not do so with unreported, manifestly illegal images in the same photo-folder[15]. This indicates that “notice and takedown” is being implemented only in the narrowest possible manner.

2.2 Lessons learned regarding (lack of) transparency

Lesson: "notice and take down" is based on trust
Content moderation on platforms is currently a total “black-box”. The public is asked to trust companies and their so called “transparency reports”, which are a non-verifiable form of self-reporting. No other industry that generates potentially harmful services or products is allowed to only self-report and self-evaluate.

Lesson: "notice and take down" is based on chance
There is no effective, systematic and continuous monitoring of the platforms in relation to illegal content. This means that it is possible for companies to claim that they remove or block 99 % of illegal content while illegal content remains abundant on those very same platforms in large quantities[16]. Most current DSA related proposals don’t address the issue of monitoring

obligations for illegal content, be it by the companies or by third parties, in any potentially effective way.

Lesson: Transparency needs to be verifiable and explainable
The "Ethics Guidelines for Trustworthy Artificial Intelligence" of the "EU High-Level Expert Group on AI" highlight the importance of transparency and explainability of automated decision-making systems that have significant impact on people's lives[17]. Hence, an essential part of transparency is the function to explain both the technical processes of the applied automated decision making-systems and the related human decisions.

3) Recommendations for an effective regulation against terrorist content online

1) Recommendation: Mandate an actual transparent reporting and compliance system

A transparency report deserving the name reporting as well as compliance system must include explanations of the individual automated moderation tools, the technical compliance system as a whole as well as information that enable a better understanding of the application of moderation policies in practice. Such transparency would also lead to more accountability and would allow regulators to apply sanctions when appropriate. The main features of comprehensible transparency and questions to be addressed in a transparency report can be found in the CEP Policy Brief "Terrorist Content Online - How to build comprehensible transparency for automated decision-making systems (ADM)"[18].

2) Recommendation: Establish an independent and capable oversight body

An independent and capable EU regulatory compliance body should exercise effective oversight of compliance with the applicable DSA rules, including of the contractual rights narrated in the terms of service drafted by the hosting service provider with regards to content management, auditing of algorithms for use in content moderation and curation.

3) Recommendation: Require proactive search for manifestly illegal content

Due to the extraordinary amount of user generated data, particularly on the big platforms, any "notice and take down" system that aims at protecting EU citizens against harmful illegal content requires a parallel mandatory systematic and continuous search for manifestly illegal content online. The idea that users first need to be exposed to harmful content and then have to notify for-profit companies to act, is not only hard to console with duty of EU policy makers to protect citizens from preventable harms. It also ignores that bigger companies are very capable to "police" their platforms for unwanted content, such as legal nudity based on internal, commercially motivated decisions. The protection of EU citizens from harmful illegal content cannot be left to companies, users few police taskforces as well as Internet Referral Units alone. Independent and capable civil society organizations and for-profit contractors should be

commissioned and financed accordingly by the companies, but also by state actors as part of their oversight responsibility.

4) Recommendation: Regulate technology to protect civil rights

Large scale global social media companies announce frequently that they already use automated image- and text recognition systems in the form of upload- and re-upload filters to find illegal or unwanted content.[19] The issue regarding content moderation obligations for platforms therefore is not IF (upload-)filters should be applied but HOW to apply them. Smart regulation that focuses on transparency, verifiability and effectiveness of automated systems would therefore protect civil liberties more than no or weak regulation.

5) Recommendation: Enable access to automated tools for smaller companies

A licensing agreement from big platforms for smaller platforms to use their automated decision-making systems would likely reinforce big-tech dominance. Instead, a public-private partnership, comprised of a consortium of governmental and non-government actors like complienace/oversight bodies, universities and companies, should be created to produce an automated decision-making software which would serve as a “minimum standard” - tool set, that will be available and required for all companies that are regulated by the DSA.

4) Learnings from the regulatory and compliance structures of the financial industry

Tech industry faces similar basic challenges to financial industry

The financial industry, social media and content hosting platforms as well as messenger services are experiencing enhanced risk levels of being misused by terrorist individuals and entities since both industries provide services that are central to the operational needs of terrorists. While services of the financial industry are necessary to raise, store, transfer and convert assets,[20] the services and products of social media and content hosting platforms as well as messenger services provide an opportunity for terrorist individuals and entities to communicate, coordinate, propagate, recruit and transfer skills.[21]

Both industries face similar basic challenges in defending such misuse of their services. Chiefly among them are their global reach, customer identity and data protection necessities, as well as challenges to identify a small number of problematic or illegal incidents within a significantly larger set of transactions/data.

Since both industry sectors face similar risks as well as comparable strategic challenges when defending against the misuse of their services, lessons learned from the global regulatory and compliance framework of the financial industry, which developed after 2001[22] provide valuable insights that can be used to develop a similarly effective regulatory framework for social media and content hosting platforms as well as messenger services.

Lesson: Ensure uniform minimum regulatory standards

In order to ensure that this global industry sector is able to develop robust compliance mechanisms, governments cooperated in order to establish global and regional institutions in order to develop and support the implementation of uniform minimum standards for compliance mechanisms. For example, the Financial Action Task Force (FATF)[23] and the European Banking Authority (EBA)[24] set minimum standards for combating the financing of terrorism and support their implementation by Member States through national regulatory authorities.

Lesson: Develop industry awareness and comparable and transparent operational industry standards

Due to the complexity connected with the implementation of the legally mandated compliance standards, financial institutions have developed industry led capacity standards and expertise to ensure that operational industry expertise is appropriate, comparable and transferable between different institutions as well as industry awareness of regulatory and compliance requirements remains high.[25]

The Global Internet Forum to Counter Terrorism (GIFCT)[26] and Tech Against Terrorism[27] are two industry led bodies aiming at increasing awareness and expertise on the issue of the misuse of industry services by terrorist organizations. However, while GIFCT primarily addresses larger scale industry stakeholders and Tech Against Terrorism is mandated to support smaller platforms and startups,[28] the combined membership of both organizations does not yet cover even a small percentage of companies within the sector.[29] Furthermore, while both organizations aim at highlighting the threat of the misuse of internet services by terrorists, the industry does not yet seem to employ a strategic focus on all issues relating to the potential misuse of their services by terrorist actors. For example, a study by the Counter Extremism Project in April 2020 confirmed that most of the major global platforms, including major crowdfunding platforms do not explicitly exclude financing of terrorism in their community standards.[30]

Recommendation*:*

With the DSA the European Union has the opportunity not only to establish a detailed mandatory compliance system for social media and content hosting platforms as well as messenger services against the misuse of their services by terrorist individuals and entities. New provisions within the DSA could also establish a structure, that allows governments to develop and monitor the implementation of uniform minimum compliance standards, comparable to the FATF. An oversight body could then pool the resources and investigative capabilities of Member States to develop detailed guidance documents, typology reports of terrorist behavior online as well as issue trends and risk assessments which would allow for adjusting compliance standards continuously as the tactics and strategies of terrorists develop.

Recommendation:

The DSA presents an opportunity to encourage or mandate greater transparency and inclusiveness of the industry and its associations. For example, the tech industry should develop transparent and transferable standards, as far as human content moderation is concerned.[31] This should involve the development of appropriate curricular as well as industry wide accepted certifications in cooperation with educational institutions. The also DSA also provides for an opportunity to implement the necessary building blocks of internal monitoring, reporting, supervisory and audit procedures for social media and content hosting platforms as well as messenger services. Developing greater strategic awareness within the industry as well as employing transparent and transferable industry standards would not only increase the effectiveness of the defense mechanisms deployed by the industry as a whole but would also support in particularly smaller platforms and startups in developing their respective compliance systems and capacities.

Endnotes

[1] https://ec.europa.eu/commission/presscorner/detail/en/IP_20_962
[2]https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/649326/EPRS_BRI(2020)649326_EN.pdf
[3] https://ec.europa.eu/home-affairs/news/eu-internet-forum-major-step-forward-curbing-terrorist-content-internet_en
[4]https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper%20April%202020%20ENG.pdf
[5]https://www.counterextremism.com/sites/default/files/CEP%20TCO%20ADM%20Transparency%202604.pdf
[6] https://www.theverge.com/2020/5/26/21270659/facebook-division-news-feed-algorithms
[7] https://farid.berkeley.edu/downloads/publications/arxiv20.pdf
[8] https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate
[9] https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce
[10] https://www.forbes.com/sites/jemimamcevoy/2020/08/04/study-facebook-allows-and-recommends-white-supremacist-anti-semitic-and-qanon-groups-with-thousands-of-members/
[11] Article 14 of the ECD
[12] https://edri.org/digital-service-act-document-pool/
[13]https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper%20April%202020%20ENG.pdf
[14]https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper%20April%202020%20ENG.pdf
[15]https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper%20April%202020%20ENG.pdf
[16] https://transparency.facebook.com/community-standards-enforcement#dangerous-organizations
[17] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[18]https://www.counterextremism.com/sites/default/files/CEP%20TCO%20ADM%20Transparency%202604.pdf
[19] https://about.fb.com/news/2020/05/combating-hate-and-dangerous-organizations/
[20] See for example: Financial Action Task Force (FATF), Terrorist Financing Risk Assessment Guidance, July 2019, available from: https://www.fatf-gafi.org/media/fatf/documents/reports/Terrorist-Financing-Risk-Assessment-Guidance.pdf
[21] See for example: Brian Fishmann, Crossroads: Counter-terrorism and the Internet, Texas National Security Review, Vol. 2 Issue 2 February 2019, available from: https://tnsr.org/2019/02/crossroads-counter-terrorism-and-the-internet/

[22] Efforts to combat the financing of terrorism gained significant traction after the al-Qaida attacks in New York and Washington D.C. on 9 September 2001. This resulted in the United Nations Security Council resolution 1373 (2001), decided on 28 September 2001, which obligated Member States to develop stronger defense mechanisms against the financing of terrorism. See: https://undocs.org/S/RES/1373(2001)

[23] The FATF, founded in 1989, originally focused on money laundering (AML). After 2001 the organization added a strong focus on combating the financing of terrorism (CFT). See: History of the FATF, available from: https://www.fatf-gafi.org/about/historyofthefatf/ The organization developed 40 recommendations focusing on AML/CFT focus, see: Financial Action Task Force (FATF), International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation. The FATF Recommendations, June 2019, available from: https://www.fatf-gafi.org/media/fatf/documents/recommendations/pdfs/FATF%20Recommendations%202012.pdf

[24] See for example: https://ec.europa.eu/info/sites/info/files/diagram_aml_2018.07_ok.pdf

[25] See for example: Chartered Institute for Securities and Investment, Global Financial Compliance, available from: https://www.cisi.org/cisiweb2/cisi-website/study-with-us/compliance-risk/global-financial-compliance

[26] https://www.gifct.org

[27] https://techagainstterrorism.org

[28] https://techagainstterrorism.org/project-background/

[29] According to GIFCT, the current members of the organization are: Facebook, Microsoft, Twitter, YouTube, Dropbox, Amazon, LinkedIn and WhatsApp, see: https://www.gifct.org/leadership/ Tech Against Terrorism does not publish a list of its members but only an overview of its industry partners: https://techagainstterrorism.org

[30] Counter Extremism Project, Financing of Terrorism and Social Media Platforms, April 2020, available from: https://www.counterextremism.com/sites/default/files/CEP%20Policy%20Paper_Terrorist%20Financing%20und%20Social%20Media_April%202020%20FINAL%20EDIT.pdf This study confirmed the findings of a 2019 study by the Royal United Services Institute (RUSI) as part of the GIFCT Global Research Network. See: om Keatinge, Florence Keen, Social Media and Terrorism Financing. What are the Vulnerabilities and How Could Public and Private Sector Collaborate Better? Global Research Network on Terrorism and Technology: Paper No. 10, Royal United Services Institute 2019, available from: https://rusi.org/sites/default/files/20190802_grntt_paper_10.pdf. Demonstrating that platforms had not taken action more than a year after having been made aware of this loophole in their community standards.

[31] Currently, there seems to be no openly available information on the details of what standards and expertise individual industry stakeholders deploy in human content moderation. The only available detailed information has been published through investigative journalism. See for example: Elizabeth Dwoskin, Jeanne Whalen, Regine Cabato, Content moderators at YouTube, Facebook and Twitter see the worst of the web — and suffer silently, The Washington Post, 25.07.2019, available from: https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/