Council of the European Union

Brussels, 9 December 2020
(OR. en)

12735/20

CT 100
JAI 944
COSI 196
CATS 81
DIGIT 113
CYBER 224
HYBRID 33
TELECOM 211
COMPET 544
IND 202
RECH 436
DATAPROTECT 119
ENFOPOL 295
SE 7

NOTE

From:	EU Counter-Terrorism Coordinator
To:	Delegations
Subject:	The role of algorithmic amplification in promoting violent and extremist content and its dissemination on platforms and social media

In the fight against terrorist and violent extremist content online, the EU has so far focused on illegal content identification and removal as well as on promoting counter-narratives. However, these efforts are largely undermined by the way recommendation algorithms are fine-tuned by tech companies to keep their users online for business purposes.

Widely used by most large-scale social media platforms and commercial web sites (retail, streaming platforms, social media, etc.[1]), recommendation algorithms aim at maximising watch time to generate advertisement revenues. This business model drives users automatically to content that is detected by algorithms as being the most engaging. As such, social media companies do not simply offer a platform on which citizens exercise their freedom of expression. They amplify certain content and demote other content, while shielding the mechanisms that regulate this from public scrutiny. Hence, **algorithmic amplification, as it currently exists, erodes the freedoms of expression and information.** Terrorists and violent extremists aim to destroy these freedoms, whereas governments protect them when they combat terrorism. Moreover, the freedoms of expression and information constitute a crucial instrument in the fight against hatred and radicalisation. They allow hatred and falsehoods in violent extremist propaganda to be exposed and challenged, and they are essential in ensuring public support for our counter-terrorism policies and legislation, which underpins their legitimacy.

Despite important improvements achieved in recent years, online platforms too often remain a **conduit for polarisation and radicalisation**. In practice, recommender algorithms frequently promote types of content linked to strong negative emotions, including extremist and divisive content, undermining the visibility of more nuanced content (including counter-narratives challenging violent and extremist views). Some companies are reported to promote what they call 'borderline content', that is legal content which is close to the content they ban on their platforms but is most likely to keep users online; most of this content could be considered as legal but harmful (see Annex I). As a consequence, **not only illegal content can be amplified** before its detection and take-down, or in the event of reappearance, but the **automated amplification of legal but potentially harmful content, especially conspiracy theories, may implant the seed of polarisation, bring some people to embrace violent extremism and terrorist propaganda or even turn to violence**.

[1] Although all major players use recommendation algorithms to boost sales (Amazon, Netflix, etc.), this note will focus more on Facebook and YouTube (owned by Google). YouTube has one of the most extensive and sophisticated recommendation systems, and its ecosystem amplifies extremist content within mainstream discourse.

Since recommendation algorithms are continuously being developed, problems could become even more worrying in the future[2]. A lot has already been done by the EU on the removal of terrorist content online, but algorithmic amplification plays an important role which needs to be fully addressed in this specific context, as well as in the wider context of digital services regulation.

I. What's wrong with algorithmic amplification practices

1. Algorithmic amplification practices undermine the freedoms of information and expression, which are essential in the fight against violent extremism and terrorism and rejected by terrorist and violent extremist groups. Enjoying these freedoms in a democratic society helps counter the radicalisation of individuals or prevents them from turning to violence against people or public institutions.

Individuals have no right to amplification of their speech (no 'freedom of reach'). On the contrary, algorithmic amplification can, when it determines which content gets the most exposure, based on commercial criteria[3], first, significantly stifle the reach of some content covered by free speech and, second, have a distorting effect on users' ability to access a pluralistic set of diverging opinions, and hence no longer allows the free marketplace of ideas. By reinforcing the market share of a certain type of content and creating a *de facto* one-sided window for the user, it distorts access to nuanced, factual and diverse information.

Genuine pluralism of ideas is key to fighting against radicalisation and disinformation and, more broadly, essential for social cohesion, democracy and public institutions which are targeted by terrorist and violent extremist groups[4]**.** Addressing algorithmic amplification would make it possible to improve freedom of information and expression. This is not about content removal, but about mitigating amplification's sides effects.

[2] *Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse*, Birgit Stark and Daniel Stegmann, Algorithm Watch, 26 May 2020.
[3] Including possible paid contributions to generate advertising revenues. Already having strong popularity and influence as a content producer will reinforce the visibility and market share of your content to the detriment to others.
[4] See JRC Report '*Technology and Democracy: understanding the influence of online technologies on political behaviour and decision-making*' (2020). The democratic foundations of our societies are under pressure from the influence of the social media on our political opinions and our behaviours.

2. Recommendation algorithms are programmed to maximise the watch time of users by amplifying ever more engaging content which is often violent and antagonising. Among the main types of algorithms used on the web, recommendations and news feed / timelines algorithms are the two specifically designed to maximise consumers' watch time[5]. Driven by powerful artificial intelligence (AI), those algorithms are constantly improved to better perform the task set by the developers of retaining users' attention.

Many recommendation algorithms **prioritise the most engaging content, i.e. content detected by algorithms as able to keep users online for the longest time[6], which in practice is often violent and divisive,** feeding conflict and aggravating antagonism between users according their gender, race, community, political views, religion, etc[7]. Through the fine tuning of algorithms, platforms seek to maximise the 'curve of engagement' by **promoting what they call 'borderline content'**, content existing at the boundary between allowed and prohibited content under their terms of service (ToS), because engagement rises faster when this line is approached (see Annex I)[8]. Polarising and/or sensationalist - yet most often legal – content (violent content[9], conspiracy theories, fake news, anti-vaccine videos, flat earth theories, clashes, racist, hatred, lewd content[10], etc.) is more effective at keeping users' attention.

[5] As regards search engines, see for example the enquiry published in September 2019 on Twitter by Stop Hate Money, which shows top-ranked books with anti-Semitic and conspiracy theories by the search engines of book retailers' websites (*Fnac, Amazon… Pourquoi leurs moteurs de recherche valorisent des ouvrages complotistes*, Le Monde, 15 septembre 2019).

[6] On YouTube, content with a short duration but which leads users to stay a long *time* online afterwards will be amplified. Recommended content will become automatically 'popular content' (with a high *number* of views), even if it was not before, since 70% of views come from recommendations. On Facebook, recommendation algorithms are based on interaction (likes, shares, comments, etc.), but the effect is similar.

[7] See, for example, how the algorithms of social media platforms feed into the propensity of individuals to share false information: *What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble*, P. M. Napoli, Federal Communications Law Journal 70/2017–2018.

[8] The curve of users' engagement has been described by Facebook's CEO in a paper on 'borderline content', defined as '*sensationalist and provocative content*' (https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/). It confirmed extreme content drives more engagement on social media.

[9] To be distinguished from incitement to violence, which is unlawful (UK Online Harm White Paper, 2019).

[10] On platforms such as Instagram, pictures with nudity are better ranked in users' newsfeeds by algorithm, pushing content providers to adapt their approaches.

However, it might be possible to use the technology for the prevention of radicalisation. As an application of AI systems for improving users' experience, recommender systems could reduce the accessibility of content through the tuning of search engines or automated detection to moderate content. The machine learning technique called natural language processing (NLP) could help to contextualize language translations, classify slang or dialects used as hate speech, and decrypt coded language used by certain groups to avoid automated takedowns[11].

3. Algorithmic amplification's problems are rooted in business models based on watch time. Maximising viewer engagement to sell advertisements constitutes the business model of many big social media platforms, especially for companies such as Facebook or YouTube[12]; more than one billion hours are spent on YouTube every day, with 70% generated as a result of recommendations by its algorithm[13].

Some companies have no incentive to promote a variety of viewpoints or content that is not addictive since recommending polarising content remains the most efficient way to expand watch time and gather more data on customers, to better target advertising and increase the returns. After internal debate, Facebook opted against increasing content checks to preserve its business model based on the mastering of data on time spent, likes, shares and comments, describing the promotion of more civil conversations as 'paternalistic' and fearing accusations of political bias[14].

[11] See, for right-wing extremism, https://www.vice.com/en_us/ article/7kpm4x/the-boogaloo-bois-are-all-over-facebook
[12] YouTube uses recommendation algorithms in various places but the main metric is the time spent online after viewing content. YouTube is estimated to have US$ 15 billion in annual revenues.
[13] https://qz.com/1178125/youtubes-recommendations-drive-70-of-what-we-watch/
[14] https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499.

II. The impact of algorithmic amplification practices on CT/CVE policies

1. Amplification of illegal content

The amplification of illegal content cannot be addressed by means of removals alone. Despite remarkable progress since 2015, a lot of terrorist content remains on platforms[15], hence the terrorist content online regulation (TCO) should be quickly adopted.

In support, we should ensure illegal content has not been amplified before removal takes place: indeed, the terrorist content or illegal hate speech could have been online for some time and likely to have been considerably amplified by the algorithms, often leading to millions of views and causing far greater damage than in the absence of amplification before removal[16].

Terrorist and violent extremist groups are particularly good at weaponising algorithms[17]. Jihadist as well as right-wing terrorist and violent extremist groups have understood how content from a small pool of hyperactive users can disproportionately influence public discourse[18]. Right-wing violent extremist and terrorist groups post videos almost daily, for instance clips relating to news events, to ensure that their videos appear high in the recommendations among less extreme clips. They have also mastered the use of hashtags, catchy and optimised titles and key words, as well as the skilful use of film techniques, so that they attain top ranks in search engines and in the integrated search algorithm of the YouTube platform that puts content on the home page, creating a feedback loop.

Amplification practices attract those groups. On channels without amplification such as Telegram, their reach becomes significantly smaller[19]. **Terrorist and violent extremist groups reinforce each other in amplifying illegal content,** either by competing to make their own content more visible, or by supporting the same type of content (hatred of democratic values, anti-Semitism, etc.).

---

[15] *ISIS 'still evading detection on Facebook', report says* - BBC News, 13 July 2020 (The Institute for Strategic Dialogue tracked 288 Facebook accounts linked to a particular ISIS network over three months and found that ISIS members were able to '*exploit gaps in both the automated and manual moderation systems on Facebook*' to generate more views of ISIS material).

[16] Moreover, amplification often makes it more difficult to identify the person who posted an illegal video (form of 'information laundering').

[17] *The Virus of Hate: Far-Right Terrorism in Cyberspace*, 5 April 2020, ICT.

[18] https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

[19] See example for White Supremacists: https://www.dailydot.com/debug/yiannopoulos-complains-telegram/

Amplification of **illegal hate speech** can contribute to the spread of ideology. The EU has started to address right-wing violent extremist content online, but not yet Islamist extremist ideology, which contains a lot of illegal hate speech. Such content has contributed to the radicalisation of a number of the recent attackers, showing a continuum between illegal hate speech and terrorism[20].

2. Amplification of legal harmful content that may be conducive to radicalisation and violence

Algorithmic amplification contributes to mainstream extreme views and 'normalises' legal but harmful content. Although the 'filter bubbles' and 'echo chambers' effects, meaning a state of 'informational isolation', are debated but deserve our strong attention (see Annex II), many studies have documented the impact on radicalisation of recommender systems[21] and companies are fully aware of this. In 2016, according to an internal presentation at Facebook, 64% of all extremist group joins were due to Facebook's recommendation tools[22]. Extremist content environments can facilitate abuse of freedom of expression and of other fundamental rights through the hidden presence of illegal content amongst banter, irony and very offensive language[23].

[20] Several Member States have specialized prosecutors on hate crimes and/or hate speech, or are about to create specialized offices (in Germany, dedicated hate speech prosecutor's offices should be created after the changes in the Network Enforcement Act (Netzwerkdurchsetzungsgesetz) come into force; in France, the future specialized judicial pole for illegal hate speech will work closely with the 'Parquet national antiterroriste', the antiterrorism judicial pole).

[21] In 2018, the sociologist Zeynep Tüfekçi described YouTube as the '*Great radicalizer*' in https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html ('*YouTube may be one of the most powerful radicalizing instruments of the 21st century (...) YouTube leads viewers down a rabbit hole of extremism, while Google racks up the ad sales*'), illustrating how recommended videos about vaccines lead to antivaccine conspiracy theories, and videos about US politics lead to '*white supremacist rants, Holocaust denials and other disturbing content'*. Her article built on a study published by a former engineer at YouTube (https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478; https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth; https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/). The Pew Research Center underlined how the site's recommendation engine steers users toward progressively more extreme and popular content as measured by view counts (https://www.pewresearch.org/internet/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/).

[22] https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

[23] The shooters at the synagogue in Poway, the Walmart in El Paso and the mosque in Christchurch in 2019 all posted on 8chan before committing their terrorist attacks (*Artificial Intelligence and Countering Violent Extremism: A Primer*, Global Network on Extremism and Technology (2020)).

This overexposure to extremist content may lead to violence, including terrorism. Automated amplification of extreme fringe content may facilitate radicalisation (by driving extremists towards more extremist opinions[24] and making people in the mainstream more likely to support extremist ideas and legitimise violent extremism) and exacerbate polarisation in society[25]; some users could ultimately be driven to explore illegal content or to act violently in real life[26].

In addition, algorithmic amplification spreads mis/disinformation[27], which also supports the rise of extremism and gives credit to terrorist propaganda[28]. The Covid-19 pandemic has revealed the nexus between illegal content (hate speech and terrorism) and legal harmful content (conspiracy theories and disinformation[29]). Not only are the boundaries between these types of content sometimes blurred[30], but disinformation can work as a tool to recruit new followers for extremist ideologies, which can lead to real-life violence[31].

---

[24] For instance, some users who initially engage with relatively prevalent forms of extreme right-wing content end up commenting on the most extreme fringes of right-wing violent extremist content (https://www.tubefilter.com/2020/01/29/this-is-why-researchers-studying-radicalization-on-youtube-need-to-be-logged-in/)

[25] On how social media may indirectly contribute to polarization by facilitating a distorted picture of the climate of opinion, leading to an overrepresentation of radical viewpoints and arguments in political discourse, see *Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse*, Algorithm Watch, 26 May 2020).

[26] On causality between online hate speech and real-life violent crime, see *Fanning the flames of hate: Social media and hate crime*, Karsten Müller and Carlo Schwarz, SSRN Electronic Journal, 2019. See how misogynist ideology, as seen in the Incel movement, could fuel terrorism (*Artificial Intelligence and Countering Violent Extremism: A Primer*, GNET (2020)). '*Antisemitic conspiracy myths are often the initial step that may lead to hatred, hate speech, incitement to acts of violence and hate crime*' (Council Declaration on mainstreaming the fight against antisemitism across policy areas, 2 December 2020, doc. 13637/20).

[27] Disinformation is intentional (to deceive, cause public harm or make economic gain), whereas misinformation is not (JOIN(2020) 8 final of 10.6.2020 *Tackling COVID-19 disinformation - Getting the facts right*). See also definition by the EU Code of Practice on Disinformation. The European Democracy Action Plan (COM(2020) 790 final of 3.12.2020) includes information on influence operations and foreign interference in the information space. The note will use 'disinformation' to cover the whole spectrum.

[28] Especially on YouTube (*Radical Filter Bubbles - Social Media Personalisation Algorithms and Extremist Content*, Global Research Network on Terrorism and Technology, Paper n°8. YouTube, compared with the Reddit platform, is specifically prioritizing extreme right-wing violent extremist material after interaction with similar content). There could even be a link between the development of online platforms and the rise of right-wing violent extremist movements across the EU and the US (https://www.investigate-europe.eu/publications/disinformation-machine/; https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html: '*YouTube is not just a driver of radicalization; it is a full-fledged far-right propaganda machine*').

[29] On how it can drive people down '*algorithmic rabbit holes*' to conspiracy theories or white supremacist propaganda, see (https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430 and https://www.nytimes.com/2018/09/07/world/europe/youtube-far-right-extremism.html). Algorithms make social media particularly vulnerable to disinformation.

[30] JOIN(2020) 8 final of 10.6.2020 *Tackling COVID-19 disinformation - Getting the facts right*.

[31] Studies confirm that a stronger conspiracy mentality leads to increased violent extremist intentions; see *Conspiracy Beliefs and Violent Extremist Intentions: The Contingent Effects of Self-efficacy, Self-control and Law-related Morality*, Bettina Rottweiler, University College London (2020).

Terrorist and violent extremist groups depend on hate speech to radicalise and recruit followers; they thrive in a climate of suspicion and distrust of political institutions[32]. Terrorist groups actively propagate their own conspiracy theories[33]; some conspiracy theories deny the reality of terrorist attacks[34].

Disinformation is likely to fuel the emergence of new forms of violent activism or terrorism rooted in conspiracy theories[35], including technophobia[36], by impacting how individuals place themselves in relation to democracy and the government and, more generally, in relation to the mainstream[37]. We have recently witnessed a huge disinformation campaign carried out by Third States in some Member States to create the impression that Europe is 'Islamophobic', which is a concept used by Islamist extremists to silence debate and all criticism, and deny the role of ideology in violent radicalization[38].

[32] According to a study, 'radical violent extremists' (RVE) groups are significantly more likely than 'non-violent extremists' and 'moderates' groups to use conspiracy theories and promote violence, both for groups focused on radical Islamic fundamentalism as well as white supremacy groups (*The Truth is Out There: The Prevalence of Conspiracy Theory Use by Radical Violent Extremist Organizations*, Terrorism and Political Violence, Rousis, Gregory et al., 19 November 2020).

[33] The purported will of Western countries to destroy Islam is the baseline of Jihadist propaganda; right-wing extremists and terrorists claim that a small elite facilitates the replacement of white people by migrants.

[34] https://www.theguardian.com/us-news/2017/oct/04/las-vegas-shooting-youtube-hoax-conspiracy-theories

[35] The famous 'Pizzagate' clearly shows how potentially deadly online conspiracy theories can become in the offline world (https://www.nytimes.com/2017/06/22/us/pizzagate-attack-sentence.html).

[36] We have already seen small-scale acts of violence caused by a belief in conspiracy theories (e.g. against 5G telecom masts: https://www.nytimes.com/2020/04/10/technology/coronavirus-5g-uk.html) and, given the amount of disinformation online, we could see more serious examples of this in the future, such as increasingly violent ecologist and animal rights groups.

[37] Echo chambers / filter bubbles can promote disinformation by ensuring that users do not see rebuttals or other sources that may disagree; they can also mean that users perceive a story to be far more widely believed than it really is (*The Disinformation Communications Challenge*, ESCN, July 2018).

[38] See the challenge of avoiding the 'trap of stigmatization' advocated by Islamist extremists (*Islamist Extremism in Europe: Challenges for Practitioners*, Magnus Ranstorp, Radicalisation Awareness Network, November 2020).

This situation is especially alarming because **the impact of amplified content on opinions is potentially massive.** With two billion active users per month[39] and over 500 hours of video content uploaded every minute, YouTube is the second biggest website and the largest video streaming platform[40], and covers approximately 95% of the internet population[41]. Consumption of recommended content on this platform is very widespread[42] and algorithm-driven content may not be perceived as 'biased'[43]. Teenagers tend to consider content posted by YouTubers as **authentic**[44].

3. Undermining of non-violent alternative content and counter-narratives. Owing to the prioritisation of extreme content by algorithms, non-violent alternative content and counter-narratives are not promoted and become invisible. Paradoxically, searching extremist content for research work or fact-checking will reinforce its overall amplification.

III. The need for recognition of the problem and more transparency by companies

1. By selecting content and directly determining what content users see, social media platforms are not neutral hosting service providers. Although some of these companies already existed when the liability exemption for hosting digital services was adopted as part of the e-commerce Directive (eCD)[45], the widespread use of recommendation algorithms has deeply changed the nature of the social media and digital providers, which are powerful actors promoting certain ideas and influencing users' opinions and, more generally, users' relationships to society and political institutions.

[39] *A longitudinal analysis of YouTube's promotion of conspiracy videos*, Marc Faddoul, Guillaume Chaslot, and Hany Farid, March 2020.
[40] https://www.alexa.com/siteinfo/youtube.com. 73% of US adults may use YouTube (https://www.pewresearch.org/internet/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/)
[41] YouTube is present in 91 countries and accessible in 80 different languages according to the company. See https://influencermarketinghub.com/social-media-statistics/#:~:text=YouTube%20has%20launched%20local%20versions,%25%20of%20the%20Internet%20population.
[42] '*some 81% of YouTube users say they at least occasionally watch the videos suggested by the platform's recommendation algorithm, including 15% who say they do this regularly'* (https://www.pewresearch.org/internet/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/).
[43] See https://www.their.tube/ which provides real-time examples of what recommended videos will look like on your YouTube home page, based on your political beliefs.
[44] https://gnet-research.org/2020/03/02/youtubes-role-as-a-platform-for-extremism/.
[45] For example, companies such as Amazon, Google or Alibaba were existing before the adoption of Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce').

Through the design of such algorithms to actively influence the visibility of different content items, platforms are not just a simple conduit for information, but they proactively select information and also proactively push information to users to make money. Platforms are active intermediaries and their automated services are not passive. The notion of 'active hosting' in the jurisprudence of the Court of Justice of the EU (CJEU) remains based on a case-by-case approach[46].

Companies already prohibit large portions of legal content that does not comply with their ToS by closely monitoring the content available on their platforms through filtering tools to detect and remove it, exerting a form of 'algorithmic content curation' for commercial reasons[47]. In some countries, they are able to make accessible only selected political content[48]. Some companies actively microtarget their users, with highly personalised advertisements being directed at users based on their personalities[49].

Therefore, the liability regime should take into account the use of amplification technologies, differentiating illegal content from legal content.

2. Recognition of the problem and more transparency by companies regarding their tools and practices is crucial. Until recently, YouTube for example argued that users were the cause of the problem (consumers are served with what they want[50]) and were benefitting at the same time from being exposed to varied content. The company argued that extreme content did not lead to a greater degree of engagement[51].

---

[46] Case C-324/09 *L'Oréal SA and Others v eBay International AG and Others*, CJUE, 12 July 2011. Having control and knowledge over data stored would prevent the operator of an online marketplace from benefiting from the liability exemption, it would be therefore considered as active but only for that *specific piece* of illegal data. The mere use of algorithms or automatic means to select, organize or present the information would not be per be sufficient to automatically meet the 'active' role standard, especially not over specific pieces of content.

[47] https://ec.europa.eu/jrc/en/news/social-media-influences-our-political-behaviour-and-puts-pressure-our-democracies-new-report-finds

[48] *Facebook touts free speech. In Vietnam, it's aiding in censorship*, Los Angeles Times, 22 Oct. 2020; *Facebook and Instagram are censoring protests against police violence in Nigeria,* Vice, Oct. 2020; *Facebook manipulated the news you see to appease Republicans, Insiders say*, Mother Jones, 21 Oct. 2020. Facebook played a 'determining role' in the violence against Rohingya Muslims, according to the UN (https://www.economist.com/asia/2020/10/22/in-myanmar-facebook-struggles-with-a-deluge-of-disinformation).

[49] Facebook's algorithm, analysing only 300 likes, can predict a user's personality with greater accuracy than their own spouse (https://ec.europa.eu/jrc/en/news/social-media-influences-our-political-behaviour-and-puts-pressure-our-democracies-new-report-finds)

[50] A study published in December 2019 that concluded that radicalisation has more to do with the people who create harmful content than the site's algorithm (*Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization,* https://arxiv.org/pdf/1912.11211.pdf) has been comprehensively criticized by a Princeton computer science professor (https://twitter.com/random_walker/status/1211262124724510721).

[51] *A longitudinal analysis of YouTube's promotion of conspiracy videos*, by Marc Faddoul, Guillaume Chaslot, and Hany Farid, March 2020; see https://www.nytimes.com/interactive/2020/03/02/technology/youtube-conspiracy-theory.html.

In addition to the major well-known US internet companies, other foreign companies delivering digital products and services in the EU are concerned by algorithmic amplification. In particular, **TikTok,** which has very quickly gained popularity among young people, is facing increasing volumes of terrorist content, xenophobic or anti-Semitic content, as well as disinformation[52]. It should also be explored whether algorithmic amplification is an issue in the context of **online gaming**.

For transparency, companies should share more information related to facts, figures, statistics, etc. concerning their amplification algorithms and their impact, at all relevant levels (national, regional and international) for public purposes but also with governments and regional/international institutions. Recommendation engines are very complex and largely opaque to the users[53], public authorities as well as non-governmental organisations.

Not all advertisers are aware of the type of content promoted. Companies share only a few specifics about their internal research[54], their effort (e.g. the setup of a Common Ground team against polarisation at Facebook, statements that the company changed their algorithm to decrease recommendations from 'borderline' and conspiracy content[55]) and the results. Closed Application Programming Interfaces (API) restrict access by independent researchers to data[56]. Such transparency related to recommendation algorithms would facilitate an informed debate in parliaments and society as well as regulatory oversight.

---

[52] See how TikTok is used by Islamic State to spread propaganda videos, BBC News, Oct. 2019 (https://www.bbc.co.uk/news/technology-50138740). The platform was hosting more than 80 million views of TikTok videos with PizzaGate-related hashtags in June 2020 (*A TikTok Twist on 'PizzaGate'*, the NYT, 29 June 2020). See *Far-Right Activists Are Taking Their Message To Gen Z On TikTok*, HuffPost, 16 April 2019; https://scramnews.com/tiktok-investigates-britain-first-tommy-robinson-far-right-hate-videos/

[53] The algorithms are so complex that even their developers have sometimes a hard time explaining them (conversation with Guillaume Chaslot).

[54] Internal research conducted at Facebook in 2016 revealed that their algorithms were responsible for the growth of extremist groups (https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499).

[55] YouTube's leaders have said repeatedly that they were addressing the content problem. In April 2019, the company indicated that it was starting a new approach by adding tracking 'quality watch time' to the time tracking but it did not share whether it had abandoned 'watch time' (https://fortune.com/2019/04/11/youtube-metrics-quality-watch-time/).

[56] By restricting access to data via the API (which is a software intermediary that allows two applications or software products to communicate by exchanging data and functionalities- see Art. 2 (18) of Directive (EU) 2018/1972 of 11 December 2018 establishing the European Electronic Communications Code) to only selected people, companies hinder independent researchers from extracting significant amounts of data to build empirical work. Hence, it allows companies to argue that the criticism emanating from research is based on too narrow a data set for it to be representative, whereas the independence of the conclusions from the 'allowed' research could be called into question (see http://thepoliticsofsystems.net/2016/05/closing-apis-and-the-public-scrutiny-of-very-large-online-platforms/).

3. Companies' policies for legal harmful content could be improved and made more consistent. The line between allowed and prohibited (legal) content varies across platforms; some companies make borderline content subject to limited functionality by, for example, removing it from recommendations and search results on the platform. Besides, **companies struggle with effective enforcement of their own policies.** Companies can be effective in not amplifying illegal and legal harmful content[57] as well as in dealing with virality[58] but results vary: although the number of conspiracy theory videos YouTube recommends[59] decreased after a backlash from advertisers in January 2019, one year later clips denying climate change continue to flourish on YouTube[60]. Facebook has been a huge conduit for conspiracy theories about the COVID-19 pandemic[61]. Extremist content creators are capable of circumventing the company's rules[62].

4. Platforms move sometimes when public pressure increases but this is not enough. In early April 2020, YouTube announced it would suppress content promoting 5G coronavirus conspiracies because several 5G masts had been set on fire in the UK[63]. Facebook shifted its policy regarding content moderation after important advertisers started to boycott the platform in the context of racial tensions in the US[64]. However, over time, market forces (withdrawal of advertising, bad publicity, etc.) are mostly not able to push companies to really change their recommendation practices[65], because private businesses benefit from the wide audience of social media for their advertisements.

[57] For example, YouTube inserted a text link that brings users to public health information pages relating to COVID-19 to fight dis/misinformation (https://www.theguardian.com/world/2020/apr/05/youtube-to-suppress-content-spreading-coronavirus-5g-conspiracy-theory).
[58] For example, after the bombings of churches and hotels in Sri Lanka at Easter in 2019, Facebook prevented the resharing of posts by friends of friends, to stop inflammatory content travelling too far or fast (https://www.economist.com/briefing/2020/10/22/social-medias-struggle-with-self-censorship).
[59] Including those claiming that the US government helped organise 9/11, or that the earth is flat (*A longitudinal analysis of YouTube's promotion of conspiracy videos*, Marc Faddoul, Guillaume Chaslot, and Hany Farid, March 2020; https://www.nytimes.com/interactive/2020/03/02/technology/youtube-conspiracy-theory.html).
[60] https://www.technologyreview.com/2020/03/03/905565/youtube-halved-conspiracy-theory-videos-recommends·
[61] https://www.wsj.com/articles/coronavirus-misinformation-spreads-on-facebook-watchdog-says-11587436159?mod=article_inline
[62] Some conspiracy theorists are using collaborations and interviews as a workaround, getting other YouTubers to either host them or talk about them on their channels (*How covid-19 conspiracy theorists are exploiting YouTube culture*, MIT Technology Review, 7 May 2020). See previous footnote on the use of coded language to avoid detection.
[63] https://www.theverge.com/2020/4/5/21208956/youtube-suppress-false-5g-coronavirus-conspiracy.
[64] This campaign is known as '*Stop Hate for profit*'; it was initiated by the Anti-Defamation League and the National Association for the Advancement of Colored People.
[65] For instance, important change in Facebook's policies was expected after its CEO published an op-ed just after the Christchurch attack (https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html).

IV. The EU is tackling online harm through sectorial regulation and various frameworks

1. The Audiovisual Media Services Directive (AVMSD)'s update constitutes an important step for content regulation but remains limited to a certain type of providers and content, and does not change liability exemptions. The AVMSD requires national legislation to ensure that video-sharing platform providers[66] under their jurisdiction take appropriate measures to protect minors from harmful content (such as gratuitous violence and pornography - see Annex I) and the general public from illegal content such as incitement to violence or hatred as well as public provocation to commit a terrorist offence, or offenses concerning child pornography, racism and xenophobia.

Contrary to the eCD, the AVMSD has certain extra-territorial scope, so as to capture services that purposely avoided their establishment in the EU[67]. The Commission has put in place a specific framework with Member States for its implementation[68], also engaging with video-sharing platforms and civil society stakeholders on the specific issue of media literacy tools.

However, the duty of care established by the AVMSD as regards illegal content and legal harmful content only applies to video-sharing platform providers and not to other hosting service providers; measures as regards harmful content only cover content which may impair the development of minors, and do not cover other types of harmful content such as disinformation. In surplus, the AVMSD refers to the liability regime set out in the eCD[69], which means that the AVMSD does not change the fact that video-sharing platforms benefit from those liability exemptions. Currently, the AVMSD has not yet been fully transposed by Member States[70].

---

[66] A video-sharing platform (VSP) is a subcategory within information society services and a certain type of Hosting Service Provider (HSP). A VSP can be a full service or a dissociable section of such service. For example, YouTube is both an HSP and a full VSP, whereas Facebook would have segments of the service which are only a HSP (hence eCD applies) and some which are also a VSP (both eCD and AVMSD apply).

[67] In addition to services established in a MS in accordance to the eCD, the AVMSD has extended its territorial scope to services not established themselves in the EU but which would be deemed to be established when a parent, subsidiary or any other entity of the group is established in a Member State.

[68] The Commission engages with ERGA (European Regulators Group for Audiovisual Media Services, composed of national independent media regulators) and the Contact Committee, composed of Ministries. A report on the application of the AVMSD is published regularly, but there is, at this stage, no meaningful data on the implementation of the new rules on the protection against incitement to violence, hatred and terrorism on video-sharing platforms because those rules were introduced in the last revision adopted in November 2018 and to be transposed by MS by September 2020.

[69] The video-sharing platforms as defined by the AVMSD do not have editorial responsibility over the content they host, contrary to audiovisual media services. All the measures established by the AVMSD are without prejudice to articles 12-15 of the eCD, hence video-sharing platform providers have to comply with both the eCD rules under Article 14 as well as the relevant AVMSD provisions.

[70] On 23 November 2020, the Commission opened infringement procedures against 23 Member States for failing to transpose the revised Directive on audiovisual content (AVMSD).

2. With regard to online content, the EU has developed a framework, **leading three separate multi-stakeholder processes** to *voluntarily* detect and fight illegal or harmful material online (see Annex I). First, the EU Internet Forum (EUIF), led by DG HOME, deals with terrorist, extremist and child sexual abuse content[71].

Second, in 2016 the EU adopted a Code of Conduct on countering illegal hate speech online[72], where monitoring reports are discussed within the High-Level Group on combating racism, xenophobia and other forms of intolerance led by DG JUST. Third, the Code of Practice on Disinformation, published in July 2018[73], is implemented through a Multistakeholder Forum on Disinformation led by DG CONNECT[74], complimentary to the Action Plan against Disinformation[75], which deals with legal harmful content.

Even if those frameworks tackle a different type of content, alongside the respective legal definitions, most of the stakeholders are the same non-European big tech companies covering most of the Internet and digital services[76]. In addition, those initiatives share a lot of common issues, such as definitions of content, the role of technology versus humans, the manipulation of content[77] or the role of online gaming industries[78], as well as policy challenges such as transparency, challenging companies' self-assessment, appropriate monitoring, technological developments, media literacy and users' awareness, best practice sharing, and empowering civil society and the research community.

[71] In the field of strategic communication towards terrorist and violent extremist groups, one could mention the role of the European Strategic Communication Network (ESCN) and the Radicalisation Awareness Network (RAN).
[72] The commitments of the signatories are monitored on the basis of the total number of notifications of content deemed to be 'illegal hate speech' sent by organisations located in Member States, as well as self-assessment reports emanating from the companies; the main metric is the removal of notified content.
[73] COM(2018) 236 final of 26.4.2018 *Tackling online disinformation: a European Approach*.
[74] Not to mention the EEAS EastStratcom Taskforce and the Rapid Alert System, aimed at monitoring networks for external interference, the European Cooperation Network on Elections, which monitors the influence on our electoral integrity, or the funding of the European Digital Media Observatory.
[75] JOIN(2018) 36 final of 5.12.2018 *Action Plan against Disinformation.*
[76] Google (YouTube), Facebook, Twitter and Microsoft hosted consumer services (e.g. Xbox gaming services or LinkedIn), Instagram, Google+, Dailymotion, Snap and Jeuxvideo.com cover 96% of the EU market share of online platforms that may be affected by hateful content (JHA Council 7 October 2019 - *Information note: Progress on combating hate speech online through the EU Code of conduct 2016-2019*).
[77] SWD(2020) 180 final of 10.9.2020 *Assessment of the Code of Practice on Disinformation - Achievements and areas for further improvement*. In particular on bot-driven amplification or the involvement of influencers.
[78] See EU CTC doc 9066/20 *Online gaming in the context of the fight against terrorism* (6 July 2020).

Moreover, the main focus so far has been on detection and removal mechanisms, as well as on the promotion of counter-narratives, with ex-post monitoring. It is a great move that the next meeting of the EUIF in January 2021 will discuss algorithmic amplification. But, whereas the impact of recommendation algorithms is well known in other areas such as consumer protection[79], it would be important to include this subject in all work strands.

V. Possible way forward: recommendations for rectifying the harmful effects of algorithmic amplification

1. Reinforce knowledge and awareness of algorithmic amplification

(1) Foster public and private research. The EU could boost research capabilities to better understand the phenomenon of amplification (including cross-platform algorithmic amplification[80], if and how those algorithms reinforce themselves and potentially interact with search engines) as well as all types of algorithms impacting the visibility of content (downgrading, upgrading). The EU should enhance research on the nexus between terrorism, hate speech and legal harmful content such as disinformation and the impact of algorithmic amplification in this context[81].

Since YouTube's algorithms have become a focus of increasing research interest, the EU should support data science and anthropological as well as sociological research to better understand patterns of human behaviour online and 'reverse engineer' the functioning of algorithms and their observable impact on users (e.g. radicalisation phenomenon, the conduciveness to violence of already radicalised people, protective factors for violent extremism[82], etc.).

[79] For example, for the propagation of hidden advertising but also scams through paid contributions. The recent Directive on Better Enforcement and Modernisation of Consumer law requires transparency in such contributions.
[80] In the context of terrorist content dissemination, Europol has witnessed the widespread use of bots to also spread content across platforms on the Internet.
[81] In France, the 'Plan national de prévention de la radicalisation' seeks to support applied research to fight against 'l'enfermement algorithmique'. Mesure 14 : '*Soutenir les travaux de recherche appliquée sur les processus d'enfermement algorithmique. Contribuer au développement d'outils pour sortir de l'exposition à des contenus susceptibles d'encourager une dérive radicale et promouvoir efficacement le contre-discours*.'
[82] Such as strong law-relevant morality: see *Conspiracy Beliefs and Violent Extremist Intentions: The Contingent Effects of Self-efficacy, Self-control and Law-related Morality*, Bettina Rottweiler, University College London (2020).

Beyond amplification, **the EU should obtain a complete picture of the business models of the attention economy[83] and the software power of persuasion**, including the interplay between suggestion algorithms, bots[84], influencers, ads[85], paid content[86] and virality. In particular, the abuse of bots by terrorist/violent extremist organisations have become a very concerning trend to evade auto-detection systems of platforms and recolonize messaging applications or social media platforms[87]. It would also be important to study how **social media ecosystems may lead to contamination of mainstream content by violent extremist content** (see Annex III). Additionally, it might be useful to measure how companies could financially benefit from amplification of illegal and legal harmful content. It would also be important to keep up with the pace of innovation with regard to algorithms and bots, the development of new forms of social media[88], as well as evolutions in users' practices in sharing content[89].

The EU could support independent research and foster its own capacities. It could further mobilise the Commission's Joint Research Centre (JRC) on algorithmic recommendation[90], as well as incorporate inputs from the Internet Referral Unit (IRU) at Europol with regard to terrorist content. Some bridges could be built with US research centres[91].

---

[83] *Artificial Intelligence and Countering Violent Extremism: A Primer*, GNET (2020).
[84] Bots are a computer program that automates interactions with web properties over the Internet. Although not 'recommendation algorithms' per se, bots have become a major tool for artificially amplifying content (they operate accounts, boost the apparent popularity of accounts by followers bots), or pushing down some content by overwhelming social media feeds.
[85] For example, it would be interesting to assess if mainstream content could be interrupted by paid-for extremist ads.
[86] With financial contributions (even of a limited value), developers can also favour specific content and generate a loop of amplification. See findings from NATO's Centre of Excellence in Riga (*Falling Behind: How social media companies are failing to combat inauthentic behaviour onlin*e, Bay, S. & Fredheim, R. NATO STRATCOM COE, 2019). https://www.stratcomcoe.org/how-social-mediacompanies-are-failing-combat-inauthentic-behaviour-online).
[87] While benevolent bots are widely used by social media to automate services (e.g. index or moderate content, feed news, detect copyright law violations, etc.), malicious bots, disguised as human users and behaving in an either partially or fully autonomous fashion, are used on a large scale to access a server, network, or web proper and run their programme. They allow to disseminate widely lists of URLs pointing at terrorist content on a number of different platforms, thanks to specific functions, such as getting customised notifications and news, follow instructions, and interact with users. These bots can be bought and sold on the black market. See, for example, *Hateful People or Hateful Bots? Detection and Characterization of Bots Spreading Religious Hatred in Arabic Social Media* (https://arxiv.org/pdf/1908.00153.pdf).
[88] Google Docs, where anyone can view and anyone can edit anonymously, has become a social media during the coronavirus and protest against the police brutality in the US (*How Google Docs became the social media of the resistance*, MIT Technology Review, 6 June 2020). Youbo, launched in 2015 to create video livestreams with up to 10 friends, has reached 40 million users worldwide (https://www.forbes.com/sites/igorbosilkovski/2020/04/18/yubo-social-platform-for-teens-triples-its-daily-new-users-amid-the-coronavirus-crisis/).
[89] For example, on how to handle the challenge raised by private communications, such as closed messaging groups that increasingly tend to replace open public debate and become a leading conduit for spreading disinformation.
[90] The JRC Report '*Technology and Democracy: understanding the influence of online technologies on political behaviour and decision-making*' (2020) exposes the role of 'algorithmic content curation' where algorithms prioritise content that has, or is expected to have, a high level of engagement, creating the risk of an overexposure to polarising content and underexposure to less emotive, but more informative, content.
[91] Such as the Stanford Cyber Policy Center or the Harvard Law School. See recent discussions in the US, for example, around Art. 230 of the Communications Decency Act and practices of companies such as Twitter labelling fake news.

(2) Promote user awareness of algorithm amplification and content 'contamination' through social media ecosystems. Initiatives such as https://algotransparency.org/, which exposes daily recommended videos on YouTube, could be promoted[92]; information on algorithmic amplification should be part of EU initiatives aimed at developing digital literacy and consumer awareness campaigns[93]. This could include supporting campaigns, such as that led by the Mozilla Foundation called 'YouTube regrets', to share testimonies from users on 'bad' content to which they have been exposed[94]. These campaigns could target young people, who use platforms and social media in large numbers and could potentially be more influenced by harmful online content.

2. Enhance EU's engagement with companies.

(3) The EU should engage more with companies in CT/CVE frameworks. First, the EU should increase its engagement with companies through **the EU Internet Forum (EUIF)**. The topic of algorithmic amplification will be put on the agenda of the next EUIF. Considering the importance and transversal nature of the topic, the EUIF would be reinforced by the systematic participation of all relevant Commissioners. A strong participation of Ministers of the Member States is important.

[92] https://www.technologyreview.com/s/610760/an-ex-google-engineer-is-scraping-youtube-to-pop-our-filter-bubbles/
[93] For example, the #ThinkBeforeSharing campaign with UNESCO to stop sharing conspiracy theories and disinformation. Work in the fields of media literacy and informed decision-making should build on the European Democracy Action Plan and the 2020 EU Citizenship Report.
[94] https://foundation.mozilla.org/fr/campaigns/youtube-regrets/

Some objectives could be established, such as creating common standards for the responsible design and application of algorithmic recommendations within the EU and at global level[95], or stimulating companies to foster innovation aimed at avoiding the amplification of illegal content and borderline content[96]. Companies could be encouraged to amplify counter-narratives such as testimonies of the victims of terrorism or violent extremism, especially to counterbalance amplified content denying the reality of terrorist attacks, as well as those of disengaged terrorists/extremists[97]. The EU could encourage companies to start sharing more information related to the amplification algorithms and their impact immediately; with the input of experts, a blueprint with specific data categories could be developed. In addition, the EUIF should also address the dissemination of illegal content through social media interaction.

Second, **the EU should raise the issue in the Global Internet Forum to Counter Terrorism (GIFCT)**. The work accomplished to mitigate the virality of terrorist content through a crisis protocol with the EU was an important step. The EU, in coordination with Member States, should seek to actively influence the agenda by bringing the work done within the EUIF inside the governmental body as well as within the different working groups, especially those focusing on algorithm outcomes and research, so as to provide common metrics and share knowledge and good practices.

(4) Streamline related activities on terrorist content, hate speech and disinformation. First, the EU should explore how algorithmic amplification can be best integrated into hate speech and disinformation dialogues with the internet companies.

[95] This work could build upon non-binding requirements on transparency suggested by the High-level Expert Group on AI as well as the proposals of the White Paper on AI. Especially on preserving human autonomy and avoid adverse effects thanks to human oversight, traceability of algorithms, accountability, privacy and data governance, robustness and safety, accuracy, diversity, non-discrimination and fairness, or the promotion of societal well-being. AI systems need to be developed in a responsible manner and with a proper ex-ante and ex-post consideration of the risks that they may generate.

[96] Solutions to ensure that smaller companies and startups have access to the appropriate technology should be considered, such as public-private partnerships and license agreements (see for example https://www.counterextremism.com/sites/default/files/CEP%20Policy%20Paper_EU%20DSA_Sept%202020.pdf)

[97] Tools such as Google Redirect are useful, but it is only focused on advertisement in search results.

Second, it may be useful to **better coordinate or even merge the different EU work streams and dialogue fora with the companies** focusing on terrorist content online, illegal hate speech and disinformation, so as to tackle the interaction between those contents, through conspiracy theories notably. **At the very least, the two frameworks on illegal content could be merged** and better coordinated with the framework for disinformation, where important work is ongoing[98]. The EU Anti-racism Action Plan 2020-2025[99], the EU Action Plan on Human Rights and Democracy 2020-2024[100] as well as the European Democracy Action Plan[101] should play a role here. This new architecture should be closely **coordinated with the framework established for the implementation of the AVMSD**.

Moreover, this new EU framework should include vigilance regarding content related to violent extremism and terrorism which is amplified by third States to weaken European societies. It might worth exploring the benefits of creating a link with EU initiatives in the field of **Hybrid Threats**[102].

Such streamlining would offset the limited scope of each dialogue and **strengthen the EU in its relations with the companies.** Non-EU companies are taking advantage of the fragmentation of our response. Overall, this could contribute to the EU's wider ambition to develop **global standards** for the internet in a thriving digital society. It would also be useful, as many of these companies are based in the US[103], to engage a dialogue with US authorities on adapting the rules on online content moderation, as well as with like-minded countries[104].

---

[98] See examples of best practices in https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54455

[99] COM(2020) 565 final of 18.9.2020 *A Union of equality: EU anti-racism action plan 2020-2025*. It describes disinformation and conspiracies targeting minority communities; the role of algorithmic amplification could be developed in addition to the risk of bias and discrimination built in AI systems.

[100] JOIN(2020) 5 final of 25.3.2020 *EU Action Plan on Human Rights and Democracy 2020-2024*.

[101] COM(2020) 790 final of 3.12.2020 *On the European democracy action plan.* See in particular the objectives of developing appropriate measures to limit the artificial amplification of disinformation campaigns, reducing the monetisation of disinformation linked to sponsored content, as well as ensuring an effective data disclosure for research on disinformation.

[102] COM(2020) 605 final of 24.7.2020 *On the EU Security Union Strategy*. The Strategy support the need to mainstream hybrid considerations into policy making. See JOIN(2016) 18 final of 6.4.2016 *Joint Framework on countering hybrid threats a European Union response*. This framework seeks to build resilience to counter violent extremism and radicalisation (fight against terrorist and violent extremism online propaganda as well as against disinformation seeking to radicalise individuals, destabilise society and control the political narrative). In particular, some foreign countries strongly contribute to strengthening extremist views and divisive content online, so as to weaken trust in democracy and public institutions.

[103] Action taken by companies to prevent virality of harmful content in the context of the latest US Presidential elections shows that there are solutions to rein in algorithms.

[104] See Council conclusions on European Union – United States relations of 7 December 2020 (doc 13724/20) and JOIN(2020) 22 final of 2.12.2020 *A new EU-US agenda for global change*.

3. Ensure correct the functioning of the internal market for recommender systems

(5) Ensure a level playing field as regards recommender systems. Looking at the nearly monopolistic position of some companies on their market, competition law should be fully applied with regard to practices in algorithmic amplification (checking for instance if the use by default of in-house recommendation algorithms could be a tying practice), for instance when reviewing the role of information gatekeepers by systemic platforms in the future Digital Markets Act (DMA)[105]; this work should build on consumer protection policies, such as the Directive on Better Enforcement and Modernisation of Consumer Law, which requires transparency with regard to paid advertisements in the results of search queries[106].

(6) Develop economic incentives to avoid the side effects of algorithmic amplification. There is a need for increased competition in the dissemination of created content as well as in the ways algorithms promote certain types of content. If free technical solutions exist to deactivate recommendations (such as the Chrome extension Distraction Free YouTube), the EU could promote the emergence of European alternatives for recommendation engines, with the support of like-minded advertisers, based on encouraging amplification of less aggressive or negative emotions, valuing positive human potential instead of feeding addiction to polarising content[107]. It could be worth exploring potential support to initiatives from civil society to introduce a grading/labelling system for channels / content that would be based on the 'quality' of the content delivered[108].

[105] COM(2020) 67 final of 19.2.2020 *Shaping Europe's Digital Future*. '*Some platforms have acquired significant scale, which effectively allows them to act as private gatekeepers to markets, customers and information. We must ensure that the systemic role of certain on- line platforms and the market power they acquire will not put in danger the fairness and openness of our markets*'.
[106] According to Directive (EU) 2019/216: '*Providing search results in response to a consumer's online search query without clearly disclosing any paid advertisement or payment specifically for achieving higher ranking of products within the search results*' is always unfair.
[107] Such measures could build on the EU Media and Audiovisual Action Plan (COM(2020) 784 final of 3.12.2020 *Europe's Media in the Digital Decade: An Action Plan to Support Recovery and Transformation*).
[108] See for instance https://www.respectzone.org, which is an NGO partnering with the European Commission.

4. Digital Services Act: Enhance transparency, accountability and oversight

The forthcoming Digital Services Act (DSA) presents an opportunity to comprehensively regulate all types of digital services providers[109], content and consumers. It should update the liability regime, provide greater transparency for the industry, and to put in place the building blocks of monitoring, reporting, supervisory and audit procedures. As for other industries, especially financial services[110], accountability and compliance should be better developed to protect the public interest.

(7) Set up a detailed and compulsory transparency framework. An essential part of transparency is to explain both the technical processes of the applied automated decision making-systems and the related human decisions. The 'Ethics Guidelines for Trustworthy Artificial Intelligence' of the High-Level Expert Group on AI highlight the importance of the transparency and understandability of automated decision-making systems that have a significant impact on people's lives.[111]

In particular, companies' **transparency reports** should include detailed information concerning their practices on recommendation, whether blocked or removed illegal and borderline content was promoted by the platform's algorithms, including the number of views, as well as data on how often the content was recommended to users, and whether human oversight was involved[112]. The role of **third parties** in programming and managing services such as bots on a platform should be explained too[113].

[109] With a territorial scope as large as in the AVMSD.
[110] See, regarding disinformation, *From safe harbour to sectoral regulation: Deploying financial services regulatory theory to address disinformation in content recommender systems*, Owen Bennett, 2020.
[111] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[112] For example, include precise metrics such as the size and popularity of recommended legal harmful content such as borderline content, the means and results in automated detection of banned legal content / borderline content, share their risks assessment of legal harmful content, their content moderation tools, ad placements, etc.
[113] In particular, companies should explain the roles of the bots and third-party applications on their platforms and how they serve content to users, including detailing the management and control performed by external party.

Companies should be obliged to share raw data with Member States, EU authorities and third parties (research, NGOs)[114]. Platforms argue that the technical complexity of the functioning of algorithms places limits on the sharing of data whereas their business data are extremely precise[115]. Without hard data, yearly reports will remain unverifiable snapshots; the data would help the EU to build an understanding of recommendation processes and their impact, support research in this field, have a more informed dialogue with the companies, and improve the evidence-base for policies on radicalisation or disinformation, amongst others[116].

Companies should be obliged to share information about their own research work on recommendation algorithms. This would help bridge the information gap between researchers inside companies and 'outsiders', while fostering critical research and the detection of problems, as well as solutions that the platforms themselves did not identify.

As transparency and accurate information are a pre-condition for informed debate, as well as for regulatory oversight, policy measures and enforcement, there needs to be a legal obligation. It is important to set out specifics about the information sought while remaining open to technological developments. Voluntary engagements with regard to transparency, as suggested in relation to the EUIF, can only be an immediate first step but cannot replace legislation, as we noticed with regard to the removal of terrorist content online.

(8) Liability for recommending illegal content should become the norm, with serious financial consequences. Recommending illegal content should be prohibited by the DSA, and companies should be financially liable for harm caused by such recommendations: they would risk fines proportionate to the amplification, imposed by a newly created EU supervisor, as in the case of editors in traditional media with regard to illegal content, or similar to the fines established by the General Data Protection Regulation (GDPR).

[114] These data should be aggregated, pseudonymised or anonymised. Competitors should not have access to APIs.
[115] Cf. criticism raised by the Social Science One project initiated by Facebook which allows scholarly usage of only a limited collection of pre-defined datasets (*Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse*, Algorithm Watch, 26 May 2020).
[116] This should be carried out with due respect for intellectual property law and the protection of companies' business models.

Regulation should consider the use of automated recommender systems as an active way of exerting control over content by companies. These systems are specifically designed to target users not just to organise content in general. Given the control, there should be no exemption from liability.

Such liability would arise even in the event of takedown of illegal content, if the amplification has taken place prior to it or after reappearance, depending on the scale of the amplification. The risk of serious fines might encourage companies to be more effective in detecting, analysing and removing all illegal content, not only manifestly illegal content that has been notified to them[117], so as to mitigate the risk that items of suspiciously illegal content that have been amplified might be classified as illegal by competent authorities.

Regarding the subsequent risks of 'over-removal' of content, large companies are already able to police their platform and remove for example content that violates copyright rules, they also remove legal but unwanted content for commercial purposes, with a view to minimising the risk of pressure from the public and advertisers. Mitigating mechanisms could be designed to avoid significant restrictions in the types of speech that can be expressed[118]. A specific regime for small platforms could be set up.

[117] Results from 'stress tests' conducted on the 'Notice and Take Down' mechanism of the German Netzworkenforcement Act (NetzDG) are not encouraging (some companies took no, or only very delayed, action as regards manifestly illegal notified content or only removed the manifestly illegal content that was notified), see https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper%20April%202020%20ENG.pdf and http://www.jugendschutz.net/fileadmin/download/pdf/bericht2019.pdf.

[118] Like public-private partnerships for handling de-risking practices by banks to avoid the risk of non-compliance with Anti-Money Laundering/Counter Financing of Terrorism legislation.

(9) Create a duty of care as regards the amplification of legal harmful content for all digital services providers. Tackling the amplification of illegal content cannot be isolated from that of legal harmful content[119]. As explained above, the amplification of harmful content, albeit legal, may ultimately lead to violence and the growing interplay between those types of content calls for bolder approaches. **The objective is not to remove legal harmful content**, but rather to correct the way that content is served and consumed by mitigating its overwhelming visibility through amplification (it will remain accessible but not dominant).

The DSA could thus introduce a **duty of care** for companies to assess and address the prevalence of legal harmful content, especially through its amplification. Regulation could include general principles such as an obligation to work with civil society, cooperate with public authorities and comply with guidance emanating from national and European regulators. This could help to create the basis for a shared understanding of legal harmful content, as well as harmonising companies' ToS on banned and borderline content within the digital single market (as a kind of rulebook covering scope, measures to detect such content and neutralise its amplification without removal, including through human oversight, etc.) while ensuring that the **EU's values and norms** are respected[120].

Provisions could go further by banning the monetization of such content through its amplification. Proportionate sanctions could be envisaged for failure to comply with the duty of care, and a specific regime for non-large platforms could be designed.

Such a regime would build upon existing and/or future legislation on artificial intelligence, in order to protect fundamental rights (to prevent breaches in relation to human dignity, non-discrimination based on gender, racial or ethnic origin, religion or belief, disability, age or sexual orientation - not only privacy and the protection of personal data) or the safety of the consumer[121].

---

[119] See, for example, the European Parliament Resolution of 15 June 2017 on online platforms (2016/2274(INI)), urging platforms '*to strengthen measures to tackle illegal and harmful content*', while calling on the Commission to present proposals to address these issues. The UK Online Harm White Paper proposes the creation of an enforceable 'duty of care' for both illegal and harmful activity.

[120] Until recently TikTok's guidelines banned criticism of systems of government and the 'distortion' of historical events including the massacre near Tiananmen Square (https://www.economist.com/briefing/2020/10/22/social-medias-struggle-with-self-censorship).

[121] COM(2020) 65 final of 19.2.2020 *White Paper On Artificial Intelligence - A European approach to excellence and trust*.

(10) Set up an oversight mechanism at EU level to monitor and enforce obligations

Companies cannot be their own regulator and supervisor. As for financial services, an independent EU supervisory authority should be established to ensure that companies fulfil their obligations, to monitor practices and measures taken by companies, enforce prohibitions and apply sanctions when necessary.

Transparency needs to be verifiable and explainable. The transparency framework and in-depth technical dialogue with companies would enable the EU supervisory authority to assess the correctness of transparency reports, exercise oversight of companies' policies as well as compliance with their own ToS as regards content management (and with applicable future DSA rules) and impact. The EU supervisory authority would have the ability to conduct audits in live mode[122]. Supervision would make it possible to fully understand how recommendation algorithms work and how companies regularly tweak them, to discuss the effectiveness of their policies and mitigate their side effects[123]. A specific regime for small platforms could be designed.

The supervisory role could be assigned to the authority that may be set up to oversee the implementation of the DSA, in close relation with Member States. An effective supervision requires sufficient staff capacities as well as adequate competences (including data scientists and coders).

[122] The White Paper on AI proposes an ex-ante assessment, a continuous monitoring system and ex-post controls on high-risk AI systems, in particular through testing, auditing or certification, based on access to data or documentation, and the verification of actions or decisions that may have been taken by AI systems. The recommendation algorithms are constantly evolving, under the designer's programming, users' interactions and self-learning.
[123] It is worth noting that Facebook's Oversight Board, launched on 22nd October 2020 to scrutinize its moderation decisions and issue binding rulings, will not be able to consider posts that have been algorithmically demoted, as opposed to deleted (https://www.economist.com/briefing/2020/10/22/social-medias-struggle-with-self-censorship).

ANNEX I

Legal harmful content and borderline content – EU frameworks on online content

1. Illegal content and legal harmful content

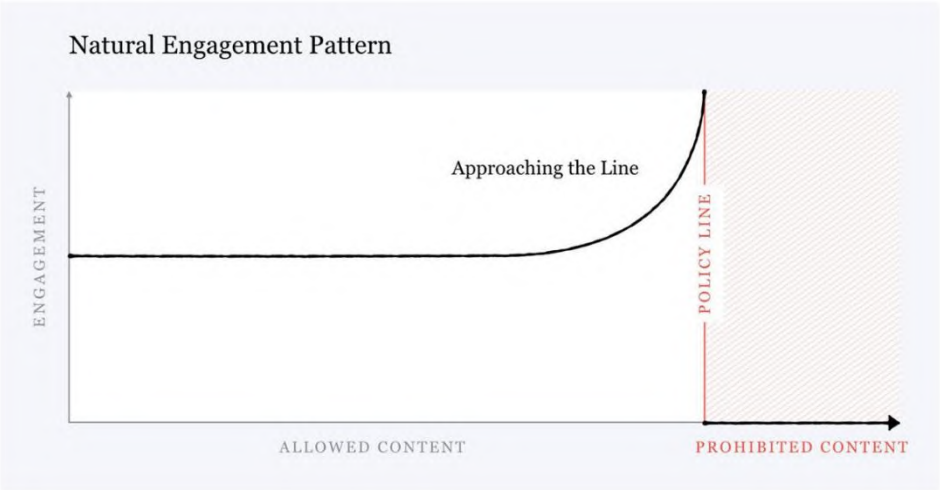
According to the European Commission, **illegal content online** means '*any information which is not in compliance with Union law or the law of a Member State concerned*' [124]. Incitement to terrorism[125], xenophobic and racist speech that publicly incites hatred and violence[126], as well as child sexual abuse material or infringements of consumer protection laws, are illegal in the EU[127].

This note refers to '**legal harmful content**' as content that does not cross the legal threshold, but which is, or could potentially be, particularly damaging to users, especially vulnerable ones such as minors, to society or democracy[128]. Legal harmful content can be subjective, depending on companies' policies or national laws, or can sometimes be legally ambiguous[129]. This content is generally protected by freedom of expression[130], which applies not only to information and ideas that are favourably received or inoffensive, but also to those which '*offend, shock or disturb'*.[131]

[124] C(2018) 1177 final of 1.3.2018 *Commission Recommendation on measures to effectively tackle illegal content online*.
[125] The 2017 Directive on Combating Terrorism defines terrorist offences but not terrorist content. The TCO is aimed at providing an agreed definition. Violent extremism is not defined by European law.
[126] Illegal hate speech is defined in EU law as the '*public incitement to violence or hatred directed to groups or individuals on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin'*. (Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law).
[127] COM(2017) 555 final of 28.9.2017 *Tackling Illegal Content Online*.
[128] The UK Online Harm White Paper refers to legal harmful content as doing harm to individuals or threatening the UK way of life, either by undermining national security, or by reducing trust and undermining shared rights, responsibilities and opportunities to foster integration.
[129] The COM(2016) 288 final of 25.5.2016 *Online Platforms and the Digital Single Market Opportunities and Challenges for Europe* as well as the Audiovisual Media Services Directive use the terminology 'harmful content' (content '*which may impair the physical, mental or moral development of minors*' in AVMSD).
[130] COM(2018) 236 final of 26.4.2018 *Tackling online disinformation: a European Approach*.
[131] European Court of Human Rights, Handyside v. United Kingdom, App. No. 5493/72 (7 December 1976), § 49.

2. Borderline content

Companies establish their own rules, which often go further than most laws. Companies' Terms of Service (ToS) distinguish 'prohibited' from 'allowed' content. Prohibited content encompasses illegal content as well as additional legal content that violates their own rules (e.g. Facebook's ban on pornography or, more recently, Holocaust denial). 'Borderline content' is allowed content, because it complies with the ToS, but it is close to prohibited content. Borderline content would encompass legal harmful content. Facebook's CEO defined 'borderline content' as '*sensationalist and provocative content*'[132].



Source: *A blueprint for content governance and enforcement*, Mark Zuckerberg, November 2018.

Although, from a legal perspective, 'borderline content' could be understood as legal content whose legality remains unclear or ambiguous (e.g. violence vs incitement to violence), this note will stick to the definition applied by companies.

[132] https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

3. The main EU frameworks on content online

	EU Internet Forum	**Code of Conduct**	**Code of Practice**
Content	Illegal (terrorist and child sexual abuse content)	Illegal (hate speech)	Legal (disinformation)
Major companies	**Facebook, Microsoft, Twitter, Google**, **Snap** and Dropbox (2015)	**Facebook, Microsoft, Twitter** and **Google** (2016). Instagram, **Snapchat** and Dailymotion (2018); Jeuxvideo.com (2019); **TikTok** (2020)	**Facebook, Google and Twitter**, Mozilla (2018); **Microsoft** (2019); **TikTok** (2020)
Main drivers	Improve terrorist and child sexual abuse content detection and removal; promote alternative and **counter-narratives**	Adoption of rules and standards; review the majority of the content flagged within 24 hours and remove or disable access to hate speech content, if necessary; develop a network of trusted flaggers from civil society; promote independent **counter-narratives** and **educational programmes**; promote **transparency** towards users as well as vis-à-vis the general public, etc.	Reduce opportunities and incentives for disinformation; **transparency** of political advertising; fight against manipulation techniques; develop tech tools that enable users to critically assess the content they access online; engage with fact-checkers and **researchers**, support **media literacy** initiatives

Source: European Commission web site.

The Filter bubble and Echo chamber effects

The *filter bubble* effect is a term coined by Eli Pariser to describe how algorithms could filter news to fit each user's beliefs[133]. The *echo chamber* effect was first described by Cass R. Sunstein in 2001.

A Deepmind paper[134] pointed to the vicious circle between decisions made by the systems that influence users' beliefs, which in turn affect the feedback the learning system receives; this feedback loop gives rise to echo chamber and filter bubble effects, which are two sides of the same coin: *filter bubble* refers to recommender systems that select limited content to serve a single user online; *echo chamber* is the self-reinforcement of one's own beliefs and preferences through repeated exposure to a certain item or category of items, by interacting with like-minded people. A large number of studies confirm this reinforcement of extremist attitudes[135].

The existence of filter bubbles and echo chambers is subject to **considerable academic debate**. The filter bubble effect has been regularly criticised by other research work, highlighting the failure to produce evidence from data[136] and the variety of online news that users of big social media platforms are exposed to[137]. It is worth pointing out that debunking arguments rely on the absence of '*very strong evidence*' of filter bubbles (rather than no evidence) and lead to the same conclusion: even with diverse content, social media may contribute to polarization in both attitudes and usage[138].

---

[133] *The Filter Bubble: What the Internet Is Hiding from You*, 2011.
[134] *Degenerate Feedback Loops in Recommender Systems*, March 2019.
[135] *A longitudinal analysis of YouTube's promotion of conspiracy videos*, by Marc Faddoul, Guillaume Chaslot, and Hany Farid, March 2020 (https://farid.berkeley.edu/downloads/publications/arxiv20.pdf). *Artificial Intelligence and Countering Violent Extremism: A Primer*, Global Network on Extremism and Technology (2020), and *Radical Filter Bubbles*, GNRTT paper n°8, RUSI (2019).
[136] *The truth behind filter bubbles: Bursting some myths*, Reuter Institute, January 2020.
[137] *How social network sites and other online intermediaries increase exposure to news*, PNAS, January 2020 (https://www.pnas.org/content/117/6/2761). This study showed that Twitter and Facebook users were exposed to a more varied online news diet than others. However, other social networks, especially those that do not allow for sharing links, such as Instagram, were not reviewed.
[138] *Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse*, Algorithm Watch, 26 May 2020. Some studies indicate that while fears that algorithmic personalisation leads to filter bubbles and echo chambers are likely to be overstated, there is evidence for possible polarization at the ends of the political spectrum.

The social media influencers' ecosystem and time viewing

YouTube's ecosystem built on celebrity and community incentivises extremist influencers both to offer more of this content to their already-radicalised audience - which is asking for more radical content from them, and to try to continuously expand their community, especially among young viewers, through trust-building, personal storytelling and apparent authenticity[139].

Social media platforms are also designed to **force influencers into a constant competition with other influencers**, which incentivises all influencers, including those producing extremist material, to constantly produce new material or material that has increased viewing time/numbers of viewers[140]. YouTube monetizes these interactions between influencers and their followers (e.g. livestreamed events such as the Super Chat introduced in 2017)[141].

Violent extremist content - especially white supremacist and xenophobic ideas - is sometimes espoused by highly visible, well-followed personalities, as well as their audiences[142]; when **appearing alongside more mainstream creators**, violent extremist narratives become disseminated throughout wider broadcasting communities, driving new audiences to their channel.

The culture of influencers has even led unwitting or unprepared personalities with a large number of followers to host campaigners with increasingly extreme views in the name of intellectual debate, curiosity or controversy[143]. The influence of groups on individual opinions is important[144].

---

[139] *Alternative Influence: Broadcasting the Reactionary Right on YouTube*, Rebecca Lewis, September 2018. The report describes an 'alternative influence network' of about 65 scholars, media pundits and internet celebrities promoting a range of right-wing political positions, from mainstream libertarianism and conservatism to overt white nationalism, broadly united by their 'reactionary' position, and presenting themselves as an underdog alternative to the mainstream media (https://www.theguardian.com/media/2018/sep/18/report-youtubes-alternative-influence-network-breeds-rightwing-radicalisation).

[140] See https://www.theguardian.com/us-news/2019/jan/08/instagram-influencers-psychology-social-media-anxiety.

[141] *YouTube Launches 'Super Chat', a Way for Creators to Make Money from Their Live Streams*, TechCrunch (blog), January 12, 2017, http://social.techcrunch.com/2017/01/12/youtube-launches-super-chat-a-way-for-creatorsto-make-money-from-their-live-streams/.

[142] https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430

[143] *How covid-19 conspiracy theorists are exploiting YouTube culture*, MIT Technology Review, 7 May 2020·

[144] See *The Law of Group Polarization* by Cass R. Sunstein, 2002, on how a group will lead to more extremist views than those of the individuals belonging to the group.