Stanford Law School

Center for Internet and Society

Crown Quadrangle
559 Nathan Abbott Way
Stanford, CA 94305-8610
Tel 650 723-1417
daphnek@law.stanford.edu

Inception Impact Assessment: Measures to further improve the effectiveness of the fight against illegal content online

Comment of Daphne Keller[1]
Stanford Center for Internet and Society
March 29 2018

Introduction

In its Recommendation on measures to effectively tackle illegal content online, the Commission proposes that Internet platforms should deploy automated content detection technologies to identify terrorist content and block or remove it. Because filters or over-zealous removal efforts may suppress lawful information and expression, the Recommendation proposes human review of algorithmically identified content, and opportunities for affected individuals to challenge ("counter-notice") removal decisions. Such corrective measures may be suspended, however, "where the illegal character of the content has already been established or where the type of content is such that contextualisation is not essential," or where content has been identified by law enforcement authorities. The Recommendation also states that, where content appears to evidence "serious criminal offences involving a threat to the life or safety of persons," platforms are to report it to law enforcement.

This Comment addresses issues unique to potentially terrorist content targeted by Internet platforms' Countering Violent Extremism (CVE) efforts.[2] It focuses in particular on Islamist extremism, though some of the analysis may be generalized to other contexts.

The Comment begins with the recognition of the grave threats posed by terrorist activity, and the acknowledged need to combat those threats, including through regulation of online content. Placing certain responsibilities on online platforms as part of this effort is appropriate. However, experience with existing platform liability regimes tells us that such legal responsibilities must be very carefully calibrated. Poorly defined and structured obligations predictably incentivize platforms to "throw out the baby with the bathwater" – silencing a

[1] Director of Intermediary Liability, Stanford Law School Center for Internet and Society; previous Associate General Counsel to Google. The Center for Internet and Society (CIS) is a public interest technology law and policy program at Stanford Law School. A list of CIS donors and funding policies is available at https://cyberlaw.stanford.edu/about-us.

[2] The discussion focuses in particular on recruitment materials or "online content constituting public provocation to commit a terrorist offence." Accordingly, it does not address the related but distinct issues of online communications that are part of the execution of terrorist attacks. Nor does it address the legally separate issue of speech inciting hatred, as distinct from inciting violent actions.

substantial margin of lawful expression beyond the genuinely unlawful content. As the Comment will explain, the resulting individual and societal harms go well beyond information and expression rights. They include pervasive discriminatory impact on Internet users based on their ethnicity, language, or religion – and they may well include real-world harms to safety and security in the face of terrorist threats.

In Section I, the Comment will review **unique attributes of potentially terrorist content**, as they affect the Commission's recommended courses of action. These include the particularly serious dangers associated with terrorist content; the complex relationship between terrorist content and lawful, important public discourse; and the role of context in distinguishing the two. It will also discuss the likely effectiveness of both filters and measures intended to correct for filtering errors, including counter-notice and human review.

The second Section will consider **discriminatory impact**. Errors in platforms' CVE content removal and police reporting will foreseeably, systematically, and unfairly burden a particular group of Internet users: those speaking Arabic, discussing Middle Eastern politics, or talking about Islam. State-mandated monitoring will, in this way, exacerbate existing inequities in notice and takedown operations. Stories of discriminatory removal impact are already all too common. In 2017, over 70 social justice organizations wrote to Facebook identifying a pattern of disparate enforcement, saying that the platform applies its rules unfairly to remove more posts from minority speakers.[3] This pattern will likely grow worse in the face of pressures such as those proposed in the Recommendation.

The third Section will focus on **security**. Improved public safety is the ultimate goal of CVE measures. It is the metric by which their success should be measured, both as a general policy matter and in balancing the interests of Internet users whose fundamental rights are curtailed. A sober assessment of the Recommendation's likely security benefits and costs is therefore imperative. This Comment cannot undertake to map out the entire security picture, which the Commission will presumably develop in consultation with experts in that field. It can, however, identify specific security costs that foreseeably arise from aggressive platform CVE enforcement. These include driving extremists into echo chambers in darker corners of the Internet; chilling important public conversations; and silencing moderate voices. Over-zealous platform removals and law enforcement reports can also build mistrust and anger among entire communities, adding fuel to existing frustrations with governments that promote such efforts, or with platforms that appear to act as state proxies. These security considerations should inform discussions of both platform monitoring and allocation of state policing resources.

Finally, Section IV will enumerate **fundamental rights concerns**. It will not closely analyze particular legal claims, but will instead list rights and foreseeable harms. In addition to the obvious concerns about information and expression rights, the Recommendation raises important concerns relating to equality and non-discrimination, data protection and privacy, and fair legal process. EU lawmakers should examine all affected rights carefully, and weigh them against the demonstrated security benefits of CVE campaigns, in determining recommendations to platforms and Member State governments.

[3] Sam Levin, "Civil rights groups urge Facebook to fix 'racially biased' moderation system," The Guardian, 18 January 2017; Tracy Jan and Elizabeth Dwoskin, "A white man called her kids the n-word. Facebook stopped her from sharing it," The Washington Post, 31 July 2017 (Facebook removing post from director of Muslim rights advocacy group documenting a threat received by a local mosque).

I. Attributes of Potentially Terrorist Content and Review Mechanisms Proposed in the Recommendation

A. Attributes of Potentially Terrorist Content

The first important attribute of potentially terrorist content is the **degree of harm** associated with it. This attribute tends to support aggressive state enforcement measures. Terrorist attacks pose extreme danger to individual safety and public order. The state's interest in preventing attacks is accordingly of the highest order. Because of the gravity of this threat, the filtering measures proposed in the Recommendation may be more likely to be necessary and proportionate, despite the burden they place on fundamental rights, than the same measures would be if used to target other kinds of unlawful content.

The second key attribute of potentially terrorist content, and one that weighs against parts of the Recommendation, is its **link to discourse on topics of public importance**. Both the causes and consequences of terrorism – including disputes over religion, immigration, regional self-determination, and more – are matters of considerable newsworthiness and legitimate public discussion. This means that true terrorist content may be difficult to distinguish from controversial or confusing, but lawful and important, expression. Platform or law enforcement errors can easily lead to suppression of important voices and public participation.

This problem makes terrorist content very different from another class of content to which it is often compared in the Internet context: child sex abuse material (CSAM). Errors in platforms' efforts to combat CSAM also affect lawful expression. But errors in the CSAM context typically involve misjudgments about the apparent age of individuals appearing in pornography. Many policymakers consider the erroneous suppression of this material to be relatively inconsequential. Whatever the merits of that assessment for CSAM, it is clearly inapposite for errors in removing potentially terrorist content. Voices accidentally silenced through over-zealous CVE efforts may be important participants in public discourse, both politically and as forces against radicalization within their communities. Their absence will distort public discussion on topics central to society today.

The third key attribute of potentially terrorist content is its **context-dependency**. This, too, weighs against depending on automation to suppress content, and in favor of robust error-correction processes. In practice, context will often be essential in determining whether a particular online communication is legal – even when a communication duplicates material previously identified as unlawful in another context. Images, video, or text concerning politically motivated violence can be illegal in one situation but important and legal in another. A standout example comes from videos posted by human rights activists to document war crimes in Syria, honor the victims, and enable future prosecution of perpetrators. YouTube has all too often taken these down, presumably because identical footage was used elsewhere by extremists.[4] Other important online information that may incorporate such content includes citizens' and civil society organizations' responses to recruitment or propaganda materials; educators' and anti-radicalization experts' critiques of those materials; and academic researchers' and news reporters' analysis. This context-dependency is another key point of difference between terrorist content and CSAM. Because the latter is illegal in every context,

[4] Malachy Browne, "YouTube Removes Videos Showing Atrocities in Syria," The New York Times, 22 August 2017; Scott Edwards, "When YouTube Removes Violent Videos, It Impedes Justice," Wired, 07 October 2017.

reliance on blunt instruments like filters poses markedly less risk of systematic error. In the terrorism context, by contrast, the risk of error is high.

B. Review Mechanisms Proposed in the Recommendation

The Recommendation's overall mechanism – automated filtering and police reporting for terrorist content, paired with human review and counter-notice in some but not all cases – is poorly calibrated to protect against removal of lawful and important online information. As will be discussed in Sections II-IV, this poses significant risks for social equality, safety and security, and for fundamental rights.

1. Filters

Technical filters cannot assess context or tell whether potentially terrorist content is actually illegal. No existing machine – be it a simple filter or the most advanced artificial intelligence – can review new material, or look at old material in a new context, and say with certainty whether it violates the law. Commercially available language-based filters, for example, miss sarcasm and jokes, and perform poorly in languages not spoken by their developers.[5] YouTube's industry-leading ContentID -- which had cost the company a reported $60 million as of 2014 and is widely believed to have cost several times that in years since – routinely generates noteworthy errors.[6] The suspension of accounts documenting atrocities in Syria, discussed above, is one example. In another, musician Ariana Grande's benefit concert for victims of terrorist attacks in the UK disappeared midstream from the artist's own YouTube account.[7]

2. Human Review

The Recommendation rightly identifies human review as an essential corrective for machine-based purges of online information. But it would eliminate such review for content previously deemed illegal – even when it appears in a new context. Even where human review is deployed, though, its effectiveness will likely be limited. Existing, human-administered notice and takedown systems consistently err on the side of removing information in the face of legal risk or complexity.[8] Empirical work on platform content removal documents significant problems even within systems that rely entirely on human evaluation.[9] There is no reason to expect that humans acting as backstops to filters will perform any better, or adequately correct for machines' mistakes. And once human errors or biases feed into a filter's algorithm, they will be amplified and applied to ever more online information.

[5] See Center for Democracy and Technology, *Mixed Messages? The Limits of Automated Social Media Content Analysis* (2017) at 14, 19 (accuracy rates in the 70-80 percent range for commercially available natural language processing filters); http://www.engine.is/the-limits-of-filtering.

[6] Jennifer Urban, Joe Karaganis, and Brianna Schofield, *Notice and Takedown in Everyday Practice* (2016) at 64 (reporting public figure of $60 million and private estimates several times higher); YouTube, How Content ID Works.

[7] Mike Masnick, "YouTube Takes Down Ariana Grande's Manchester Benefit Concert On Copyright Grounds," TechDirt, 07 June 2017.

[8] http://cyberlaw.stanford.edu/blog/2015/10/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws.

[9] Urban et al.

3. Counter-Notice

The Recommendation lists other possible means to correct erroneous removals, such as counter-notice. It and other procedural protections identified in civil society-endorsed guidelines like the Manila Principles are important harm-mitigation measures.[10] But counter-notice is clearly insufficient to offset the damage arising from filtering errors. Data on counter-notice for copyright suggests that it compensates for only a small fraction of excessive removals.[11] Internet users may be particularly unlikely to challenge wrongful removals if, as is particularly likely in the CVE context, they face language barriers or are concerned about immigration status or police attention to themselves or their families.

More fundamentally, counter-notice only protects the rights and interests of the individual who posted online material. Where the larger public interest lies in *access* to that material – by journalists, policy-makers, law enforcement, or the larger public –that interest cannot be adequately protected through procedural rights granted to a single person. This is particularly so when the individual who posts controversial content – such as a witness who records and posts footage of political violence or extremist activity – is a vulnerable bystander in dangerous and chaotic circumstances. The public interest in material of this sort can only be protected by more robust mechanisms. A starting point would be broad, public transparency that enables concerned civil society organizations and experts to crowdsource the work of finding and correcting removal errors.

4. Different Standards for Small Platforms

While recognizing the need for some corrective measures, such as human review and counter-notice, the Recommendation creates lopsided rules for small platforms and their users. It proposes pooling technologies and extending a filtering mandate to those platforms, but not extending any means of error correction. This would exacerbate existing advantages of larger platforms in content removal practices.

Research shows that small platforms' removal practices differ from their larger competitors', even under the US Digital Millennium Copyright Act (DMCA) – which includes relatively strong procedural guidance for all platforms. The most comprehensive empirical research on point found that some traditional platforms simply honored 100% of requests, and a majority took users' content down "even when they [were] uncertain about the strength of the underlying claim."[12] If companies that lack the legal and financial resources of their larger competitors are unable to scrutinize every legal request that arises now, the situation will only grow worse when they are faced with the larger number of "notices" generated by algorithms scanning all user communications. Start-ups and other smaller entities cannot, like Facebook or YouTube, hire thousands of moderators to compensate for machines' mistakes.

A legal regime in which users inevitably suffer more improper and unremedied removals on small platforms will have anti-competitive consequences. Large, incumbent platforms, with better ability to avoid and correct for such errors, will have a clear advantage. Forcing small platforms into this position – taking down too much user material, forfeiting user trust, and

[10] https://www.manilaprinciples.org

[11] http://cyberlaw.stanford.edu/blog/2017/10/counter-notice-does-not-fix-over-removal-online-speech

[12] Urban et al (2016) (Smaller platforms also described feeling "left aside in policy debates and news accounts skewed by attention to the relatively few" larger actors).

being unable to introduce the corrective measures of their larger competitors – will not improve the Internet economy or encourage future competition and innovation. And it will amplify the other problems discussed in this Comment, with consequences that burden fundamental rights and may ultimately weigh against the security interests that motivate the Commission's Recommendation.

II. Discriminatory Impact and Harm to Vulnerable Minority Groups

Reports of over-removal from platform CVE campaigns are now common. As discussed above, YouTube has taken down an unknown number of videos uploaded by human rights groups or individuals to document Syrian atrocities. In a similar vein, Facebook deleted the page of an anti-violence Chechen independence organization, and removed posts documenting Rohingya ethnic cleansing in Myanmar, reportedly because the platform had designated the minority Muslim group as extremist militants.[13]

These few newsworthy episodes are not isolated. Ordinary users, too, find that records of their communications or public participation have disappeared. One British Muslim woman, who agreed to share her story anonymously, found that a prayer she posted on Facebook had been removed. It read, in Arabic, "God, before the end of this holy day forgive our sins, bless us and our loved ones in this life and the afterlife with your mercy, almighty." Anecdotally, Koran excerpts and clerical teachings are said to be particularly frequent targets of improper removal requests, from both private entities and governments around the world.

Combatting genuine terrorist content online is of course an important goal. But doing so at the cost of silencing innocent users sharing religious materials -- and reporting them to the police for doing so -- is deeply troubling. When the law prescribes short removal deadlines and strict legal consequences for failure to identify and remove extremist material, we should expect errors of this sort to become even more common. Already vulnerable groups, including speakers of Kurdish, Chechen, Farsi, Indonesian, and other languages common in Muslim-majority regions, will bear the brunt of the harm. Troubling as their experiences are individually, they are worse when considered as a pervasive and discriminatory pattern, prompted by government mandate or pressure, effecting participation on the major communications platforms of our age.

Human harms arising from this pattern are entirely foreseeable. Researchers have found that Internet users throughout the world self-censor when they are aware of potentially being monitored – including by avoiding searches on sensitive health topics like eating disorders or depression.[14] De facto targeting based on ethnicity, religion, or language makes the problem worse. The implicit societal message from governments urging platforms to over-zealous removal through un-meetable monitoring goals is a harsh one: that members of certain ethnic or religious groups are not trusted to discuss religion or current events unsupervised, and that their information and expression rights are valued less than those of their fellow citizens.

[13] Julia Carrie Wong, "Facebook blocks Chechnya activist page in latest case of wrongful censorship," The Guardian, 6 June 2017; Betsy Woodruff, "Facebook Silences Rohingya Reports of Ethnic Cleansing," Daily Beast, 18 September 2017; "Facebook bans 'dangerous' Rohingya militant group," The Hindu, 21 September 2017.

[14] Alex Marthews and Catherine E. Tucker, *Government Surveillance and Internet Search Behavior* (2017); see also Pen America, *Chilling Effects* (2013) (journalists report avoiding writing about terrorism); Jon Penney, *Chilling Effects: Online Surveillance and Wikipedia Use* (2016).

The Commission should look closely at this discriminatory impact in weighing the proposals in the Recommendation.

III. Security Impact

Experts are increasingly casting doubt on the efficacy of broad platform content purges as a means to increase safety and security. A comprehensive 2017 literature review, conducted by London's International Centre for the Study of Radicalisation, found that only a minority of published research supported "hard" approaches such as "the restriction of Internet content for security purposes," and that "[m]ost work on this topic regards such measures as impractical at best and dangerous at worst."[15] As Centre Director and Kings College London Professor Peter Neumann put it elsewhere, "approaches that are aimed at reducing the supply of violent extremist content on the Internet are neither feasible nor desirable."[16]

Security researchers' conclusions arise, typically, from empirical study of real-world radicalization mechanisms. They are buttressed by experience with platform content removal practices outside the CVE context.

A. Radicalization and Social Marginalization

As EU Counter-Terrorism Coordinator Gilles de Kerchove put it, enhanced security depends to a substantial degree on rectifying "the sense of social marginalisation which plagues Muslim communities across Europe."[17] The Hague-based International Center for Counter-Terrorism, similarly, identified "alienation and social exclusion felt in Europe" as a trigger for radicalization.[18] One significant impact of CVE campaigns may be to increase this very marginalization.

By imposing costs not only on extremists but on the individuals and communities around them, poorly calibrated CVE efforts can reinforce the very problems they were meant to correct. Some of the key risk factors for radicalization -- feelings of alienation, exclusion, frustration, and

[15] Alexander Meleagrou-Hitchens and Nick Kaderbhai, International Centre for the Study of Radicalisation, King's College London, Research Perspectives on Radicalization: A Literature Review, 2006-2016 (2017) at 53, 56. Research favoring more content removal includes Martyn Frampton, Ali Fisher and Nico Prucha, *The New Net War* (2017) and materials published by Mark Wallace.

[16] Neumann, Peter R., *Options and Strategies for Countering Online Radicalization in the United States*, Studies in Conflict & Terrorism (January 2013) 431-459 at 437. *See also* J.M. Berger and Jonathon Morgan, *The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter*, Brookings Project on U.S. Relations with the Islamic World Analysis Paper No. 20 (March 2015) at 54 (discussing "unintended consequences of [social media] suspension campaigns and their attendant trade-offs"); Ines von Behr, Anaïs Reding, Charlie Edwards, and Luke Gribbon, *Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism*, Rand Europe (2013) (finding empirical correlations between Internet extremist content and individual radicalization, but no documented causal connection that could help identify effective interventions).

[17] Gilles de Kerchove, Amazon Editorial Review of Peter Neumann, *Radicalized: New Jihadists and the Threat to the West*.

[18] Quoted in United Nations Office of Counter-Terrorism, Enhancing the Understanding of the Foreign Terrorist Fighters Phenomenon in Syria (2017) at 14.

moral outrage -- are also foreseeable consequences of over-zealous platform removal efforts.[19] Even content removals in less personal or political realms, like copyright, often leave Internet users feeling outraged and powerless. More deeply felt indignation and outrage are to be expected among those who find their online presence, or that of friends, respected community leaders, or news sources, erased from important public forums – and, perhaps worse, reported for police investigation.

Counter-radicalization campaigners and Islamist terrorist recruiters share an important target constituency: disaffected, Internet-savvy Muslims, including immigrants and children of immigrants in places like Brussels or Paris.[20] The harm and offense of over-zealous content removal will fall on these same individuals. So, too, will the chill on public participation and sense of state-sanctioned bias and exclusion brought on by platform police reporting. People who might otherwise have spoken on controversial topics – such as causes of and responses to terrorism – will remain silent, particularly if they fear unfair treatment at the hands of law enforcement. Accepting this human cost is not merely discriminatory and disrespectful. It may also be seriously misguided as a matter of security policy.

B. The Public Sphere, Political Dialog, and Counterspeech

The Internet and social media facilitate threats, but are also key sites of what Habermas called the public sphere – a place of productive and evolving discourse, situated between the private realm and public authority. Disruption in this sphere – whether through elimination of individual speakers or of collective trust – can damage both society and security. For individuals at risk of recruitment, it may mean losing some of the most important voices opposing radicalization: those of peers and community members. "[O]rganic social pressures that could lead to deradicalization" are an important part of open, public platforms like Twitter.[21] Some studies suggest that the most effective actors in reducing online aggression are respected members of a speaker's own social group.[22] Aggressive CVE campaigns will threaten and sometimes eliminate such voices.

Importantly, the public sphere includes students, thinkers, and intellectuals from across the political and social spectrum. In the CVE context, this includes individuals who may appear "radical" to those unfamiliar with a particular region or political or religious context. Pressure on platforms is all too likely to silence the very individuals who share experiences and grievances with potential extremists – but who oppose violence and act as voices of moderation

[19] Peter Neumann, *Options and Strategies for Countering Online Radicalization in the United States*, Studies in Conflict and Terrorism vol. 36:6, 431 at 435 (2013) (citing Sageman), available at http://www.tandfonline.com/doi/pdf/10.1080/1057610X.2013.784568; Meleagrou-Hitchens and Kaderbhai at 14, https://www.wired.com/2017/06/theresa-may-internet-terrorism/, Maeghin Alarid, "Recruitment and Radicalization: The Role of Social Media and New Technology," in *Impunity*, published by DOD Center for Complex Operations ("Radicalization is more widespread where conditions of inequality and political frustration prevail").

[20] Alarid (discussing ISIS recruiters "specifically targeting those who are young and computer savvy").

[21] Berger and Morgan at 3. See also Neumann at 437 (discussing counterspeech); Alarid (negative social media posts about ISIS can be "an effective tool in counterradicalization efforts."); Benesch (speech believed to be correlated to violence during Kenyan election over-represented in closed Facebook discussion compared to Twitter).

[22] Meleagrou-Hitchens and Kaderbhai at 6; Munger, K. Polit Behav (2017) 39: 629. https://doi.org/10.1007/s11109-016-9373-5

within their own political spectrum. Accepting this disruption of the public sphere risks harmful consequences for discourse and security, both.

C. Echo Chambers

Researchers report that an important tactic of extremist recruiters is to shift conversations with potential recruits out of the public eye.[23] Ironically, CVE campaigns may have precisely the same effect. Exclusion from mainstream platforms like Twitter can drive those at risk of radicalization into increasingly concentrated and insular groups on other, smaller platforms. This "much louder echo chamber," away from more moderate community members, can "speed and intensify the radicalization process."[24] This, too, is a security cost of CVE campaigns, to be weighed carefully against expected benefits.

D. Law Enforcement Tools and Priorities

For law enforcement agencies, the Recommendation also poses complex issues about allocation of resources – between policing online information and policing offline activity. Expert opinion diverges on this issue to some extent, but the 2017 literature review identified an emerging "consensus that the Internet alone is not generally a cause of radicalisation, but can act as a facilitator and catalyser of an individual's trajectory towards violent political acts," when paired with off-line, real-world contacts.[25] In the words of a German government reporter, "the internet does not replace the real world influences but reinforces them."[26] An over-emphasis on online activity may "lead policymakers in the wrong direction when it comes to counter-radicalization programs."[27]

Allocation of resources between online and offline activity is a consequential choice. Reports suggest, for example, that attackers in both Manchester and London had been identified to police by concerned friends, but that overburdened law enforcement agencies were unable to act on the information.[28] Counterterrorism scholars Peter Neumann and Shiraz Maher described a related problem in the UK government's response to the later-convicted extremist Anjem Choudary, who is said to have inspired the London Bridge attackers. Choudary had a YouTube

[23] Meleagrou-Hitchens and Kaderbhai at 7.

[24] Berger and Morgan.

[25] Meleagrou-Hitchens and Kaderbhai at 35, see also 19, 39 ("The majority of the literature takes a nuanced position that asserts the importance of online influences without negating the requirement of offline interactions"); von Behr et al (evidence "does not support the suggestion that the internet has contributed to the development of self-radicalisation" or "that the internet is replacing the need for individuals to meet in person during their radicalisation process. Instead, the evidence suggests that the internet is not a substitute for in-person meetings but, rather, complements in-person communication.").

[26] Quoted in United Nations Office of Counter-Terrorism (2017) at 39; Meleagrou-Hitchens and Kaderbhai at 35 (a "strong case for a causal connection between such materials, and violent acts perpetrated by those found to have been in possession of them, has yet to be made.")

[27] Kim Cragin, Melissa A. Bradley, Eric Robinson, Paul S. Steinberg, What Factors Cause Youth to Reject Violent Extremism? Results of an Exploratory Analysis in the West Bank, Rand (2015) at 16; von Behr et al ("Many of the policy documents and academic literature in this area focus on online content and messaging, rather than exploring how the internet is used by individuals in the process of their radicalisation.").

[28] Emily Dreyfuss, "Blaming the Internet For Terrorism Misses The Point," Wired, 06 June 2017.

channel, but "practically all of his followers were known to him personally and were recruited face to face," the researchers explained. "It is one thing for the internet companies to pull down radical propaganda. But they face an uphill battle while preachers such as Choudary have spent years spreading their message virtually unchallenged on British streets."[29]

Finally, as a practical matter, different agencies -- domestically and internationally – may have very different strategies and priorities regarding online extremist activity. When they do not coordinate, platforms can be caught in the middle, asked both to take content down and leave it up for continued surveillance.[30] Whichever answer is the right one, failures of coordination clearly waste resources that could be dedicated to effective policing. In the allocation of law enforcement resources, this cross-border and inter-agency coordination should itself be an important priority.

The Recommendation, by urging Member States to allocate finite policing resources to online content, prioritizes one theory of harm prevention over others. Because of the serious consequences of this choice, including consequences for safety and security, it is important that this allocation be strongly supported by evidence. The security risks reviewed in this Section should be weighed carefully against benefits of aggressive CVE campaigns, and due consideration should be given to measures – such as improved transparency and error-correction procedures – that can increase accuracy and decrease collateral damage to innocent individuals in platform content removal operations.[31]

IV. Fundamental Rights and State Action

The Recommendation lays out a number of actions for platforms. If these actions were taken under clear state mandate, many would raise important questions about impact on Internet users' fundamental rights under the EU Charter. If they instead remain nominally voluntary, but prompted by clear government pressure, the questions about the state's role and the relevance of fundamental rights guarantees are different, and less explored in the case law and literature. Some recent scholarship and commentary from the Council of Europe suggests analytic frameworks for assessing fundamental rights and the obligation of states in this situation.[32]

This section will briefly enumerate specific fundamental rights affected by the Recommendation. It will not undertake in-depth analysis or offer specific legal conclusions.

[29] http://www.bbc.com/news/uk-40161333?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_axiosam&stream=top-stories

[30] See, e.g., Ellen Nakashima, "Dismantling Of Saudi-CIA Website Illustrates Need for Clearer Cyberwar Policies," The Washington Post, 19 March 2010.

[31] The Manila Principles and other civil society recommendations provide a menu of possible mechanisms.

[32] Council of Europe Committee of Ministers recommendations to member states on the roles and responsibilities of internet intermediaries (2018); Aleksandra Kuczerawy, *The Power of Positive Thinking: Intermediary Liability and the Effective Enjoyment of the Right to Freedom of Expression?.* *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, vol.8 (3) , pp. 226-237 (2017); Christina Angelopoulos, Annabel Brody, Wouter Hins, Bernt Hugenholtz, Patrick Leerssen, Thomas Margoni, Tarlach McGonagle, Ot van Daalen and Joris van Hoboken, Institute for Information Law (IViR) Faculty of Law University of Amsterdam, *Study of fundamental rights limitations for online enforcement through self regulation* (2016).

A. Equality and Non-Discrimination Rights (Arts. 20-21)

Section II of this Comment discusses the existing discriminatory impact of CVE campaigns. This disproportionate harm to Internet users discussing Islam or speaking languages associated with the religion will likely expand as a result of state pressure for ever faster and more error-prone platform content review. As a legal matter, this may burden affected individuals' rights against discrimination on grounds of race, language, religion, and ethnic or social origin, as well as rights to equality before the law.

B. Privacy and Data Protection Rights (Arts. 7-8)

As the CJEU noted in the *SABAM* cases, general monitoring mandates for platforms can conflict not only with Article 15 of the eCommerce Directive, but also with fundamental rights of Internet users, including privacy rights. Monitoring users' every online utterance requires processing substantial personal data, and must be proportionate to the state's legitimate aims.[33] Particularly for those platforms that do not already assess user traffic for ad targeting purposes, the change proposed in the Recommendation would be substantial. For platforms that do already engage in content-based ad targeting, the question would be whether adding a state-mandated purpose, or expanding existing private monitoring to target new classes of content, changes the impact on data protection and privacy rights.

In any event, the addition of a duty for private platforms to notify law enforcement when they detect users sharing particular information implicates fundamental rights in a new way. It brings the platforms' delegated function closer to the kind of dragnet state surveillance addressed in the European Court of Human Rights' *Roman Zakharov v. Russia* and *Szabó and Vissy v. Hungary* cases. In the latter, the Court clarified that surveillance operations must be "strictly necessary … for the obtaining of vital intelligence in an individual operation," with legal justification for intercepting "a specific individual's communications … in each case." The pervasive monitoring and reporting described in the Recommendation seems inconsistent with this standard. It also resembles the wide-ranging and undifferentiated interference with Internet users' data rejected by the CJEU in *Digital Rights Ireland*.[34]

C. Information and Expression Rights (Art. 11)

Barriers to state-mandated monitoring of Internet communications arising from Article 11 of the Charter and Article 10 of the Convention have been discussed in previous Comments, including those submitted by the undersigned,[35] and are likely familiar to the Commission.[36] This

[33] See, e.g., GDPR Article 6.3

[34] Digital Rights Ireland, Judgement of the Court, 08 April 2014.

[35] http://cyberlaw.stanford.edu/publications/regulatory-environment-platforms-online-intermediaries-data-and-cloud-computing-and (2015).

[36] *See* Magyar Tartalomszolgáltatók Egyesülete (MTE) v. Hungary (2016) E.Ct.H.R. 82, http://www.bailii.org/eu/cases/ECHR/2016/135.html (monitoring may not be mandated in case of defamatory speech in news forum comments); Case C-70/10 Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM), 2011 E.C.R. I-11959 para. 52); and SABAM v. Netlog, (2012) 2 C.M.L.R. at para. 48; *but see* Delfi AS v. Estonia (2015) E.Ct.H.R., http://www.bailii.org/eu/cases/ECHR/2015/586.html (monitoring requirement permissible in case of unprotected hate speech in news forum comments).

Comment will not belabor them further, except to remark on three specific mechanisms addressed in the Recommendation.

First, the Recommendation describes a process of counter-notice and reinstatement when hosts erroneously remove legal expression or information. It does not, however, spell out legal immunities for hosts that honor such counter-notices. Without such legal protections, hosts will have reason to reject even valid counter-notices except in occasional unambiguous cases. One possible model for a more carefully calibrated counter-notice provision can be found in the US DMCA, which sets out both immunities and processes to remove content once it has been reinstated after counter-notice.

Second, it is ambiguous whether the Recommendation endorses the employment of Interpol or Member State police personnel to identify *illegal* content, as opposed to the broader category of content that violates a platform's Terms of Service. If it is the latter, and government resources are deployed to combat citizens' lawful online expression, new concerns under Article 11 may be engaged.

Third, while the Recommendation urges transparency efforts by platforms to enable public scrutiny of content removal operations, it has no similar provisions for government, or for the trusted notifiers who may be given special status under the Recommendation. For purposes of public understanding and accountability, such transparency – particularly on the part of government actors engaged in identification of prohibited speech – is important.

D. Fair Process Rights (Arts. 47-48)

Finally, several elements of the Recommendation may implicate Internet users' rights to fair judicial process and defense. These rights may be implicated by any system of legal enforcement that is initiated, adjudicated, and executed entirely by private companies. More specifically, though, fundamental rights issues may be raised by the role of law enforcement in identifying, without judicial review, content that is subsequently to be filtered automatically. The lack of human review for such police-identified content, as well as law enforcement power to override counter-notice proceedings, are also significant.

V. Conclusion

As the Internet evolves, reconsideration of existing rules is appropriate. The threat of terrorist violence, in particular, warrants strong responses. Lawmakers may choose to accept costs, including burdens on Internet users' fundamental rights, that would not be justifiable in other situations. Such consequential trade-offs, however, should be acknowledged and justified by sober assessment of the facts. Real-world security gains of platform CVE efforts should be weighed against real-world security harms, as well as burdens on the fundamental rights of Internet users.