RVI ADVISORY BRIEFS

Joan Barata

Beyond Limiting and Countering: How to Promote Quality Content to Prevent Violent Extremism and Terrorism on Online Platforms

Resonant Voices Initiative

Beyond Limiting and Countering: How to Promote Quality Content to Prevent Violent Extremism and Terrorism on Online Platforms

Joan Barata

Introduction

This paper analyses the policy and legal implications related to the promotion of quality online content that supports and reinforces institutional and societal efforts to prevent, counteract, and deflate radical discourses leading to violent behaviour. This analysis will focus on content disseminated via online platforms or intermediaries, and in particular on the intermediaries providing hosting services, who offer a relatively wide range of services for online storage, distribution, and sharing; social networking, collaborating and gaming; or searching and referencing[1].

The aim of this text is to go beyond the discussions around national security-based limits to speech, and even further than the mere construction of counter-narratives aimed at responding to, or deactivating the feelings and intentions triggered by certain pieces of content.

According to the Counter-Terrorism Committee established by the UN Security Council (Resolution 1353 of 2001), apart from countering terrorist narratives, it is necessary to offer effective "alternative or positive narratives", which take into account "genuine feelings of powerlessness and alienation and provide credible alternatives, especially to vulnerable young people searching for a sense of meaning in their lives"[2].

1 See the comprehensive and detailed categorisation provided by Joris van Hoboken, João Pedro Quintais, Joost Poort, Nico van Eijk, "Hosting intermediary services and illegal content online. An analysis of the scope of article 14 ECD in light of developments in the online service landscape". European Commission - DG CONNECT (Communications Networks, Content and Technology) and IViR. 2018.

2 See the Letter dated 26 April 2017 from the Chair of the Security Council Committee, established pursuant to resolution 1373 (2001) concerning counter-terrorism, addressed to the President of the Security Council (paragraph 19). See also "Preventing Radicalisation to Terrorism and Violent Extremism. Delivering counter or alternative narratives". *Radicalisation Awareness Network* (RAN). 2019.

Freedom of expression and national security

International (and regional European) human rights standards, and particularly the jurisprudence of the European Court of Human Rights (ECtHR), expressly emphasize the crucial role that the free flow of ideas and information plays in the construction and development of a fully democratic society.

This strong protection of freedom of expression can be explained by the power and influence that words – as well as images, signs or symbols – have within societies. Speech can be an instrument of political criticism, questioning values and social principles, and therefore cause major discomfort or even shock. The opinions of others and information can arouse strong feelings of rejection. In particular, those who hold public power may feel that their activities and legitimacy are called into question when sharp criticism is expressed publicly. The ECtHR emphasizes the fact that freedom of expression not only covers "information or ideas that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the state or any section of the population. Such are the demands of that pluralism, tolerance and broadmindedness without which there is no democratic society"[3].

This means that even radical and extremist speech is protected under freedom of expression clause and it can only be restricted under exceptional, necessary and proportionate circumstances.

3 Handyside v. United Kingdom. Decision of 7 December 1976. Case number 5493/72. (§ 49). The European Court of Human Rights.

As the international rapporteurs on freedom of expression have underscored:

> *"The concepts of "violent extremism" and "extremism" should not be used as the basis for restricting freedom of expression unless they are defined clearly and appropriately narrowly. Any restrictions drawing upon a CVE/PVE framework should be demonstrably necessary and proportionate to protect, in particular, the rights of others, national security or public order. The same applies whenever the concept is invoked to limit the activities of civil society, including in relation to their establishment or funding, or to impose restrictions on fundamental rights, including the right to protest."*[4]

The OSCE Representative on Freedom of the Media has also stressed that:

> *"(…) extremist activities can be subject to legal restric-tions by States when they imply the use of violence and represent a direct and imminent threat to basic constitutional pillars and, particularly, human rights, for the purpose of severe political upheaval. Mere expression of controversial and provocative political views must there-fore be respected and protected as part of pluralistic and democratic debates."*[5]

The fact that certain forms of radical or extremist speech are legal does not mean that they may not be potentially harmful or encourage processes of radicalisation that can lead to violent actions. However, it is also important to underscore, as authors like Daphne Keller have warned, that we still know remarkably little about when extremist speech (either legal or illegal) leads to violence and how to prevent that from happening[6].

4 The United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information, "Joint Declaration on Freedom of Expression and countering violent extremism". 4 May 2016.

5 OSCE Representative on Freedom of the Media, "Communiqué on the impact of laws countering ex tremism on freedom of expression and freedom of the media". 7 October 2014.

6 See Daphne Keller, "Internet Platforms. Observations on Speech, Danger and Money". *Hoover Institution.* 13 June 2018. p. 21.

Also, as Brian Fishman - who has led the effort against terrorist and hate organisations at Facebook - has bluntly expressed: "researchers cannot reliably measure how much content terrorists post online because of the confounding effect of platform countermeasures", basically because they are not able and allowed to see it[7]. The latter is not only due to the existence of legal constraints that affect the content hosted by platforms, but also and mostly as the consequence of the content moderation policies established and implemented by platforms themselves, as it will be elaborated in the next section.

Content moderation and the law

Intermediaries moderate the content they distribute. Content moderation is influenced by a series of different factors, ranging from the protection of the right to freedom of expression, the implementation of a certain business model and the avoidance of socially undesirable or harmful speech. Such interests are often intertwined and may present interesting reciprocal tensions.

The most relevant legal systems in the world (including the European Union, EU) incentivise content moderation by platforms by ensuring that they are not penalized for good faith measures against illegal or other forms of inappro-priate content. When platforms are granted immunity for the content they handle, the law is in fact incentivising the adoption and implementation of private policies regarding illegal and other types of offensive or harmful content[8].

In the United States, Section 230 of the Communications Act of 1934 (as amended by the Telecommunications Act of 1996) provides intermediaries with immunity with regards to any content moderation decision inasmuch as they are a) adopted in good faith, and b) refer to a series of types of content or, more broadly, to "objectionable" information. Section 230 has played a fundamental role in the development of the Internet as we know it. Under the protections set by US law, platforms have the incentive to operate and expand their businesses under a predictable legal regime, to moderate the content they share, and specifically to deal with certain forms of *undesirable* or harmful speech.

7 Brian Fishman, "Crossroads: counter-terrorism and the Internet", *Texas National Security Review*: Volume 2, Issue 2 (February 2019).
8 See Joan Barata, "Positive Intent Protections: Incorporating a Good Samaritan Principle in the EU Digital Services Act", *Center for Democracy and Technology*. 29 July 2020.

In the EU, Article 14 and 15 of the so-called e-Commerce Directive[9] contain relevant provisions governing content moderation. On the basis of such provisions, hosting platforms in Europe are not liable for content moderation decisions, according to several requirements and conditions. In order to retain immunity, platforms must not have actual knowledge of the illegal nature of the activities or information they facilitate. In addition to this, EU law also prohibits the imposition of general content monitoring obligations, as well as obligations to actively seek facts or circumstances indicating illegal activity. In any case, intermediaries are allowed to voluntarily adopt their own content moderation and monitoring rules and enforce them without necessarily losing the exemption mentioned.

Thus, the legal provisions in place in the US and the EU (as well as other parts of the world, like India and several countries in Latin America) protect, to different extents, the capacity of platforms to decide how to organise, prioritise, demote, or simply eliminate content. This protection is provided via liability exemptions. As has just been mentioned, such exemptions may not only shield *negative* content decisions, but also moderation policies regarding the pieces of content that are maintained and available to users.

This is part of a very relevant debate, especially since the European Commission has committed to submit a proposal for a Digital Services Act (DSA) legislative package that, among other things, will replace the provisions contained in the e-Commerce Directive[10]. This debate focuses on whether the increase in intermediaries' "editorial" control over content through moderation practices might also justify the introduction of legal modulations affecting the degree of responsibility vis-à-vis the third-party content they facilitate. Consequently, the present paper cannot focus on this important matter. However, it is important to note that the modification (either in Europe, the US or anywhere else in the world) of the current regime of platform liability applicable to moderation decisions may also directly affect the way intermediaries handle both *good* and *bad* content.

9 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.
10 Digital Services Act package. European Commission - DG CONNECT (Communications Networks, Content and Technology). 2020.

More specifically, a shift in the responsibility of online intermediaries regarding decisions on the promotion of certain pieces or types of content, may have the effect of disincentivising such internal policies.

What is content moderation?

According to James Grimmelman, content moderation consists of a series of governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse. In the same way that many of our most important offline spaces for dialogue and expression (parliaments, town halls, academic events, traditional media programmes, etc.) have moderators and rules to keep the discussions civil and productive, platforms would also tend to promote the healthiness of debates and interactions to facilitate communication among users[11]. According to Grimmelman, one of the key elements of moderation is organising content, that is to shape the flow of content from creators to users. This activity may have different modalities, including deletion or removal of content but also other types of measures such as editing, annotating or filtering, the latter consisting of deciding the specific subset of content that finally becomes available (or primarily available) to the public. Thus, filtering does not only imply demoting or making less visible certain kinds of speech (getting closer to deletion), but also promoting or visualising certain preferred types of content.

As previously mentioned, platforms adopt these decisions on the basis of a series of internal principles and standards. Examples of these moderation systems are Facebook's Community Standards[12], Twitter's Rules and Policies[13] or YouTube's Community Guidelines[14].

11 James Grimmelmann, "The Virtues of Moderation", *Yale Journal of Law & Technology:* Volume 17. Issue 1 (2015).
12 Community Standards. Facebook.
13 Twitter Rules and Policies. Twitter.
14 Community Guidelines. YouTube.

In addition to this, it is obvious that global platforms may be forced to apply local laws through local authorities as speech can usually be associated with a specific physical territory. Platforms are also being asked by relevant bodies of the international system and civil society to align their practices with international human rights law. All these normative inputs do not only affect the way social media entities design their internal rules but also their implementation by the same private bodies.

All this poses significant challenges to content regulation. States and certain civil society groups tend to ask intermediaries to make greater efforts to eradicate harmful and undesirable content, especially manifestations of hatred, disinformation, certain forms of propaganda, references to criminal acts and other similar behaviours. On the other hand, organisations dedicated to the protection and promotion of freedom of expression, international human rights organisations and even some governments have expressed their concern that global private companies often restrict or simply eliminate ideas, opinions and other content published by users on the basis of internal rules that are considered unjustified, abusive, and ambiguous. In this context, and to mention just an example, it is relevant to note that Facebook has recently established its own – yet external – body to assess the implementation of the mentioned Community Standards[15]. The creation of the so-called Facebook Oversight Board was announced in 2018, although it was not until May 2020 that the first group of members was appointed[16].

In any case, it is clear that principles and standards currently used by major global platforms to moderate content usually go beyond the legal frameworks regarding the right to freedom of expression established in many parts of the world (particularly in liberal democracies) and are generally established in the form of negative limits to speech.

15 However, the Board does not have the ability or resources to review each and every content moderation decision by Facebook. In addition, other important areas of activity of Facebook, such as its decisions regarding the business model, commercial and advertisement policies, treatment and transfer of personal data or the determination, in general terms, of its internal regulations regarding contents (Community Standards) also fall, in principle, outside the scope of the Board. Therefore, the Board will not be able so far to issue recommendations or guidelines regarding the promo-tion of specific types of content. The Charter of the Oversight Board is available online.
16 Announcing the First Members of the Oversight Board. Oversight Board. May 2020.

Therefore, and also to sum up, there are a few fundamental functions that platform content moderation currently has in common amongst the different players:

→ eliminate content that is considered to be illegal and/or harmful or objectionable,
→ improve users' experience and thus increase the size and the quality of the community (with all the associated benefits in terms of commercial exploitation),
→ alleviate legal and regulatory pressure from State bodies.

It is important to insist on the fact that the content moderation decisions that are currently at the centre of policy and regulatory discussions are those related to the willingness and capacity of platforms to get rid of *bad* speech, particularly disinformation, extremist content, and hate speech.

For example, the very last Facebook's Civil Rights Audit (July 2020), conducted at the behest and encouragement of the civil rights community and of some members of US Congress, and with Facebook's cooperation, with reference to content moderation practices, it focuses on the new policy that bans explicit praise, support and representation of white nationalism and white separatism, and a new policy that prohibits content encouraging or calling for the harassment of others. It also looks into a series of pilots to combat hate speech enforcement errors, that is to say, to avoid "false positives" and over removal of content on grounds of hate speech[17].

In line with this, academic efforts have focused on the effectiveness of moderation practices regarding the most harmful modalities of speech and their impact on the exercise of the right to freedom of expression[18].

17 Laura W. Murphy et al., Facebook's Civil Rights Audit. 8 July 2020.
18 See for example Susan Benesch, "Proposals for Improved Regulation of Harmful Online Content", *Dangerous Speech Project*. 2020.

This is still, in any case, a very important task, as according to Daphne Keller and Paddy Leerssen, as public understanding of platforms' content removal operations, even among specialised researchers, has long been limited, and this information vacuum leaves policymakers poorly equipped to respond to concerns about platforms, online speech, and democracy. Recent improvements in company disclosures may have mitigated this problem, yet a lot is still to be achieved[19].

Much less attention is paid to possible platform policies consisting of promotion of certain content types. Elettra Bietti warns that intermediaries promote content in order to maximise user engagement, addiction, behavioural targeting, and polarisation[20]. According to Bietti, in order to change this pattern, it would be necessary to treat platforms as utilities, common carriers, or essential facilities, on the basis of a new idea of infrastructure ownership and control that would give more power to users and transform the former into spaces to "enhance human interaction, social and cultural fulfilment, and political empowerment". Platforms have tried to counter this kind of argument by showing their commitment to provide users with content that is "meaningful" to them, beyond a ranking system based on what people like and click on[21].

On the other hand, Eileen Donahoe has affirmed that platform rules are a manifestation of the free expression of the platforms themselves, and therefore they should exercise their own freedom of expression to protect democracy. This also includes the power to promote, demote, label and curate content in order to protect the users' rights to seek and receive information (and not only to speak) within platforms. In addition to this, legislators and policymakers may encourage the responsible exercise of a platform's rule-making authority and enforcement in support of de-mocracy, particularly when it comes to the development of "transparency and accountability mechanisms to assess fairness in applying platforms rules and visibility into content promotion."[22]

19 Daphne Keller, Paddy Leerssen, "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation", forthcoming in: N. Persily & J. Tucker, Social Media and Democracy: The State of the Field and Prospects for Reform. *Cambridge University Press*. 2020.
20 Elettra Bietti, "Free Speech is Circular", *Berkman Klein Center Collection*. 1 June 2020.
21 See the interview with Mark Zuckerberg by Ezra Klein "Mark Zuckerberg on Facebook's hardest year, and what comes next", *Vox*. 2 April 2018.
22 Eileen Donahoe, "Internet platforms should exercise their own free expression to protect democracy", *The Hill*. 15 August 2020.

As a matter of fact, during the period of the COVID-19 pandemic, and despite criticisms derived from the intensive use of algorithmic moderation (due to lack of availability of human moderators[23]), platforms have developed a series of efforts to make available and particularly visible accurate, useful and formative content regarding the nature and evolution of the epidemic, as well as to help users protect themselves from hoaxes and disinformation campaigns[24].

This is probably the first time when global online platforms have openly engaged in campaigns to identify and promote good content and reliable sources beyond the commercial (and legitimate) interests that usually drive their content curation and promotion practices.

Moderating terrorist and extremist content

According to the latest figures provided by Facebook, 99,6% of the content actioned on grounds of terrorism (mostly related to the Islamic State, al-Qaeda, and their affiliates) was found and flagged before any user reported it[25]. This being said, it is also worth noting that big platforms already have a long record of mistaken or harmful moderation decisions in this area, inside and outside the United States[26].

In any case, there is a tendency, particularly in the European political scenario, to state that "platforms should do more". This is at least the political argument used to push new and controversial pieces of legislation such as the recent proposal in the EU regarding terrorist content online[27].

23 See for example "Facebook: Improvements in transparency reporting more urgent amid Coronavirus pandemic", *Article 19*. 12 June 2020.
24 See the initiatives by Facebook, Twitter, or YouTube.
25 Community Standards Enforcement Report. *Facebook*. August 2020.
26 See Jillian C. York, Karen Gullo, "Offline/Online Project Highlights How the Oppression Marginalized Communities Face in the Real World Follows Them Online", *Electronic Frontier Foundation*. 6 March 2018; Billy Perrigo, "These Tech Companies Managed to Eradicate ISIS Content. But They're Also Erasing Crucial Evidence of War Crimes", *Time*. 11 April 2020; "When Content Moderation Hurts", *Mozilla*, 4 May 2020.
27 Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online. European Commission - DG HOME (Migration and Home Affairs). 12 September 2018.

If it were to be adopted, this legislation would require intermediaries to detect, identify, remove or disable access to, and even prevent the re-uploading of certain pieces or types of content, as well as the expeditious removal (one hour) of content upon the reception of an order from the competent authority[28].

The existing legislation is, in any case, insufficient and unfit to deal with the complexity of matters that platforms need to handle, especially in the most immediate instances for online speech regulation. The statutory imposition of obligations to thoroughly monitor content would contravene applicable liability exemptions (also recognised by regional and international legal standards) and transform these private entities into unaccountable law enforcement agencies and seriously affect the users' exercise of the right to freedom of expression.

What is needed is a coordinated effort between tech companies, policy makers, and experts. This effort should aim, before anything, at establishing a common understanding and sharing of knowledge about the way platforms operate, the different manners in which terrorist organisations or individuals may use platforms for their purposes, the panoply of measures and instruments available to State authorities and platforms, and the implications and trade-offs derived from different possible choices.

Once again, current discussions are based on the possible options available for platforms, law enforcement agencies or the judiciary to properly assess and detect problematic content and the steps to be taken once this happens.

28 See a more detailed description of these provisions and a critical approach in Aleksandra Kuczerawy "To Monitor or Not to Monitor? The Uncertain Future of Article 15 of the E-Commerce Directive", *Balkinization*. 29 May 2019, and Joan Barata Mir "New EU proposal on the prevention of terrorist content online: an important mutation of the e-commerce intermediaries' regime", *Stanford CIS White Paper*. 12 October 2018.

In this context, the use of content matching techniques on the basis of a database of digital fingerprints or hashes is a tool that still needs to be properly studied in terms of the impact on freedom of expression and actual effectiveness as well as possible harm to vulnerable groups. In particular, the Global Internet Forum to Counter Terrorism (GIFCT)[29] was established in 2017 by Facebook, Twitter, Microsoft, and YouTube as an industry initiative "to prevent terrorists and violent extremists from exploiting digital platforms". In fact, this initiative must also be classified as an outcome of the "platforms shall do more" motto developed by Governments, particularly in Europe. The GIFCT is currently on the path to become an independently managed body. However, concerns linked to lack of transparency, vulnerability vis-à-vis pressure coming from Governments and impact on the actions of civil society groups, still make it quite a controversial instrument[30].

Considering all these elements, it is clear that engaging in and promoting counter-narratives can be a useful tool for tech companies in order to: a) prove their commitment to counterterrorism policies and alleviate legal and regulatory purposes, b) avoid the risk of accusations of curtailing legitimate speech (particularly political speech disseminated by minority communities, and c) present themselves to users as safe environments.

Platforms' engagement with counter-narratives has taken place using different modalities, including providing content or messages directly, financially supporting certain groups or initiatives, promoting the visibility of certain types of content in the context of radical groups, conversations or searches, or facilitating ad-targeting tactics for non-profit organizations[31]. In some cases, this has been the result of active engagement or collaboration between platforms and different stakeholders[32].

29 Global Internet Forum to Counter Terrorism

30 See Chloe Hadavas, "The Future of Free Speech Online May Depend on This Database", *Slate*. 13 August 2020.

31 Brian Fishman, "Crossroads: counter-terrorism and the Internet", cited in footnote 7, p. 98.

32 See the description of several examples involving the use of platforms in the report by RAN cited in footnote 2.

The effects of these initiatives have still not been properly documented, and thus may be perceived as a predominantly PR effort. Further, many rely on micro-targeting and fall short in responding to a coordinated or systematic adversarial action. In addition to these shortcomings, there are no clear and straightforward demands or directives from Governments and civil society as to how these projects must be specifically designed, which audiences to target, and a proper assessment of potential risks and benefits.

At this present stage, platforms have neither formulated successful and general internal policies regarding the dissemination of counter – or alternative narratives[33] nor have they engaged in systemic and comprehensive efforts with all relevant stakeholders in order to discuss, understand and articulate proper strategies in this area. In this sense, it is important to underscore how, for example, regular reports by major companies regarding content moderation policies and enforcement focus almost exclusively on the measures adopted regarding the restriction or elimination of illegal or objectionable content.

Conclusion: fundamental conditions to promote counter and alternative narratives on online platforms

LEGAL FRAMEWORK

The hosting and promotion of counter-narratives requires the existence of a legal framework that does not create any barrier or disincentive regarding these activities. In particular, this is linked to two main areas:

First, as it has already been mentioned, it is necessary to count on a legal regime that incentivises the capacity of platforms to organise and prioritise third-party content, without being under the pressure of provisions that associate such moderation policies with a higher degree of liability.

33 In 2016 Google renamed its incubator Google Ideas as Jigsaw with the aim, in the words of Eric Schmidt, of using technology "to tackle the toughest geopolitical challenges, from countering violent extremism to thwarting online censorship to mitigating the threats associated with digital attacks". However, this initiative is seen by many stakeholders as an unfulfilled promise. See the article by Lorenzo Franceschi-Bicchierai, "Google's Jigsaw Was Supposed to Save the Internet. Behind the Scenes, It Became a Toxic Mess", *VICE Motherboard*. 2 July 2019.

In US legislation, platforms are shielded from liability for decisions on content when it comes to both taking it down or keeping it up. In the European Union the decision to promote the publication or particularly engage in the dissemination of certain categories of content may be considered as playing "an active role of such a kind as to give it knowledge of, or control" over the hosted content, thus becoming liable in case of illegality[34]. It would be important, in this area, to move towards a regime where liability is based on stronger and clearer criteria, particularly the existence of a takedown request based on a determination established by the competent authority (preferably a court)[35]. Otherwise platforms would be disincentivised to promote, in good faith, certain categories of content in line with public policy counter and alternative narrative dissemination objectives.

Second, as stressed by the Global Network Initiative (GNI), Governments must ensure that counterterrorism laws and policies do not undermine the development and dissemination of messages by private actors that discuss, debate, or report on terrorist activities. In this sense, it is important to count on precise and balanced legislation that provides for a proper differentiation between messages that aim to incite terrorists acts and those that discuss, debate or report on them. Special protection, in this sense, is to be granted to media and journalists who comment and report on terrorists groups and activities, or even engage directly with members of such organisations (by interviewing them, for example), when the purpose is to provide a better understanding to the public of the ways these organisations operate, and of their motivations, and this does not represent any form of incitement[36].

In this context, it is also important to note that legislation and regulation must refrain from imposing obligations to speak, either in terms of obligations for certain civil society groups (which can be perceived as close to groups in risk of radicalisation) to disseminate *good* messages, or via the imposition on platforms of the duty to actively spread them.

These measures would be extremely problematic in terms of freedom of expression (as they would introduce the possibility of *forced speech*), and put in the hands of the State very sensitive powers in terms of ideological control of the public sphere.

34 See the Judgement of 12 July 2011, case C-324/09 (L'Oréal), Court of Justice of the European Union.
35 See Joan Barata, "Positive Intent Protections: Incorporating a Good Samaritan principle in the EU Digital Services Act". *Center for Democracy and Technology*. 29 July 2020.
36 Extremist Content and the ICT Sector. A Global Network Initiative Policy Brief. *GNI*. 2016. p. 4.

PLATFORMS' MODERATION POLICIES

This paper has already mentioned some of the possible measures to be considered by platforms when engaging in and supporting counter-narratives. It is important in any case that this interest is included in some way in platforms' mission statements and that it permeates their moderation policies.

It is also relevant to mention that platforms need to present themselves as "facilitators" of the work of the actual creators of content in this area, thus avoiding engaging in more aggressive strategies that would only endanger both the effectiveness of their positive content moderation policies and the independent expert work of organisations actively working on the ground.

Last, but not least, is the necessity that platforms are subject to regular and informed assessments with respect to the effectiveness and impact of the policies mentioned here. This is a topic that is still under discussion by experts (particularly regarding which narratives are the most impactful and operative in the long term), and thus platforms need to facilitate all the relevant data for a proper expert analysis as well as be ready to implement new policies or adapt the existing ones.

To conclude, it is clear that a lot of work is still to be done (both on the theoretical and practical levels) regarding the promotion of the quality content that has been the object of this paper. In any case, it is also obvious that no meaningful achievement can be made without the constructive engagement of platforms, legislative/regulatory and law enforcement bodies, as well as relevant civil society organizations.

This paper has been produced as part of the Resonant Voices Initiative in the EU, funded by the European Union's Internal Security Fund–Police.

RESONANTVOICES.INFO