

A systemic approach to hate speech and the challenges of practical application

Session Concept & Agenda & Summary

Revised version of 13 July 2022

➤ The 3 rd VSPs & Regulation Workshop: Setting the scene	1
➤ The speakers	2
➤ A selection of issues for discussion	3
➤ Summary of the Workshop	3
➤ A selection of relevant reading material	8
➤ Agenda	10

➤ The 3rd VSPs & Regulation Workshop: Setting the scene

Aim of the VSP & Regulation workshops: *These informal, non-public, ad-hoc workshops aim to facilitate practically-oriented discussions among EPRA members on the implementation of regulation (and co-regulation) of video-sharing platforms, enhanced by the inputs of selected academics.*

❖ Previous EPRA Workshops on VSPs & Regulation:

- **1st workshop:** "NRAs and VSP Regulation" on 21 October 2021; focus: *VSPs and commercial communication, VSPs and Governance*;
- **2nd workshop:** "Focus on age restrictions" (*age assurance/age verification and age/content ratings*) on 16 March 2022; in line with the key theme "Empowering & Protecting Minors" of the EPRA Work Programme for 2022.

→ To access the VSP & regulation working papers (presentations and summaries): [here](#).

❖ Recent EPRA work on online hate speech:

A thematic session on "*Living with hate speech: from apprehending to combatting*" took place last year during the [53rd EPRA meeting](#): the discussion highlighted the complex definition of hate speech and the need to identify a spectrum of severity; the paradox of technology; the benefits and limits of AI tools and the need for a better understanding of the mechanisms behind online hate speech (see [background paper](#), including a summary of the discussion).

❖ Focus & aim of the 3rd VSPs & Regulation Workshop:

Building upon the previous session around the understanding of hate speech online, this workshop will focus on the regulation of VSPs with regard to hate speech. Renowned experts will present their proposal for a systemic regulation model applied to hate speech. EPRA members will then be invited

to share their experience in combatting hate speech on VSPs and to discuss the questions raised by such a systemic approach: its practical application, its challenges and its impact on the media regulator.

➤ The speakers

- **Lorna Woods**, Professor of Internet Law at the [School of Law of the University of Essex](#) (UK)

With extensive knowledge and a long-standing experience in media policy and communications regulation, Lorna has participated in many commissioned studies on the regulation of Internet, the co-regulation in the media field and also on media pluralism and competition issues with the European Audiovisual Observatory for instance. Regularly asked to share her expertise by governmental, national and international institutions, she also contributes to the [Carnegie UK Trust Programme](#) "Tackling Online Harms" and has developed, with William Perrin, an influential public policy proposal advocating for a systemic approach for the regulation of online platforms (*see more details below*).

- **William Perrin**, Trustee, [Carnegie](#) UK

A trustee of several charities, William is a leading expert on technology policy and has helped set the national framework for diverse regulating sectors such as media – he was notably a driving force behind the creation of Ofcom. A prominent data activist, he regularly provides advice to the UK Government, as well as to some of the world's major media and technology companies, trusts and foundations. His work within Carnegie with Prof. Lorna Woods on a statutory duty of care on social media companies has informed the UK's approach to online safety and inspired the European Commission's approach to regulating online services (*see more details below*).

Focus on the Carnegie Project:

Under the aegis of the "[Tackling Online Harms](#)" Carnegie programme, Lorna Woods and William Perrin have developed a public policy proposal to improve the safety of internet services' users.

In this regard, in order to regulate online hateful speech, they promote a systemic approach rather than content-based regulation.

- **A preventive approach to reduce hate speech**

Applying a precautionary principle – *evidence of harm may be evident but not scientifically proved* -, the authors suggest adopting a preventive approach to tackle hate speech, i.e. minimise the spread and amplification of harmful content and therefore the exposure of persons to such content.

Based on the idea that everything that happens on online media result from corporate decisions and design choices, their proposal recommends acting directly on the design of online services through a *statutory duty of care*.

Built around the key components of responsibility, risk assessment, mitigation and remediation, this statutory duty of care would oblige online social media to regularly assess the impact of their services' functioning on hate speech and to design them in a way that avoids nudging users towards potentially hateful behaviours.

- **A systemic regulation applying to all online social media**

Such a "safety-by-design" obligation would apply to any online media facilitating user interaction and engagement between users and therefore, to any online media that might spread hateful content.

To ensure effective implementation and to tackle systemic issues in online social media, their guidelines call for a *strong, expert and independent regulator*.

- **An approach suitable for all online harms.**

Lorna Woods and William Perrin emphasise the potential broad scope of application of their guidelines, likely to be relevant for *various types of online harms*. For instance, recently, they were adapted by Lorna Woods and Clare McGlynn, along with Carnegie and several women rights' organisations, to publish a [Violence Against Women and Girls Code of Practice](#).

This Carnegie project and the work of Lorna Woods and William Perrin received several awards and have inspired the 2019 [Online Harms White Paper](#) and the on-going [Online Safety Bill](#) in the UK.

➤ **A selection of issues for discussion**

- *How would such a systemic approach work in practice*
- *The role that media NRAs can play in this collaborative governance*
- *First experiences of NRAs with liaising with VSPs on hate speech issues*
- *First experiences of NRAs with developing guidance on hate speech*
- *Resources (or the lack of) that VSPs put into safety & content moderation*
- *Primacy of the English language in safety & content moderation*
- *Understanding the systems: the issue of transparency and access to platforms' data*

➤ **Summary of the Workshop**

- **Presentation by Lorna Woods & William Perrin**

CONTEXT:

Several cases around the world – (e.g. on the amplification by Facebook of military propaganda following the coup in Myanmar¹) - have brought to light the potentially dangerous effects of social media, which were originally specifically designed to reach global audiences. ***What if the way social media are designed contributes to the problem of online hate speech?***

Social media are the result of choices – *with regard to their design, features and user-facing experience* - guided by the objective of maximising the shareholders' profits. Such created

¹ Note of the Secretariat: Facebook promoted pages that shared pro-military propaganda in Myanmar, even after it banned accounts linked to the military from the platform due to human rights abuses and the risk of violence, according to [a recent report by the human rights group Global Witness](#)

environment might have negative influences and consequences on content creation, dissemination and user's experience. So far, debates around online hate speech have focused on algorithms and moderation, but, according to L. Woods and W. Perrin, there is a wider possible scope of action. The idea is to **look further up the chain of content's dissemination and identify the possible points of regulatory intervention.**

BASIC PREMISE:

- The Ruggie (UN guiding) principles rely on three complementary interdependent pillars:
 - *The government's responsibility to protect against human rights abuses by third parties*
 - *The corporate responsibility to respect human rights*
 - *The need for greater access by victims to effective remedy*
- Social media are a risky industry for its users; any risky business requires **a risk-management model** (*testing the products, regular monitoring of side effects...*)

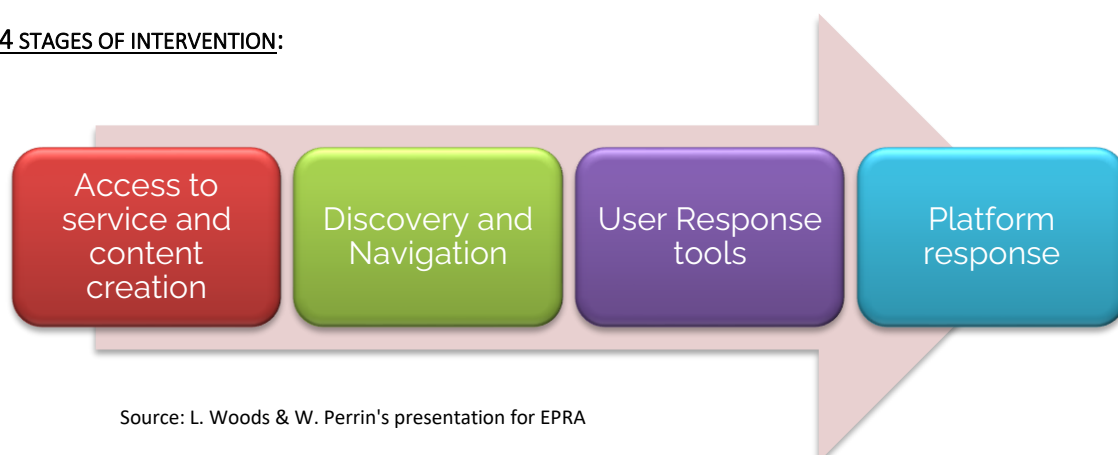
PROPOSAL:

Constraining social media to adopt a "safety-by-design" approach.

The proposal recommends acting directly on the design of online services through a **statutory duty of care**. In other words, social media providers should design their software and the related terms and conditions in a way that offers safety and positive incentives for users.

- **For providers**, it would mean implementing an outcomes-oriented regime built on a risk management model and a constant monitoring of the system for improvement.
- **For regulators**, it would mean not focusing on the identification of harmful content but on the structure of the service itself to determine what platforms should take into account when designing their services and which safe basic features should be integrated.

THE 4 STAGES OF INTERVENTION:



Source: L. Woods & W. Perrin's presentation for EPRA

Examples of design and features decisions with potential negative//positive effects on hate speech:

Access to service and content creation: The possibility to have multiple or anonymous accounts / The possibility to target audience and specific groups of people (targeted advertising) // Age assurance system recently introduced on Instagram / Ban of content related to certain topics (i.e. no vaccines discussion policy on Pinterest).

Discovery and navigation: The automatic recommender algorithms / The popularity indicators / The autoplay function (its impact on the user's opinion) // Personalisation of recommender algorithms by the user.

User response tools: The clickbait rewards (it might incite users to provide content only to receive clicks) / The metrification of the level of popularity (like, share...) // An effective complaint mechanism / Reduction of the forwarding velocity (WhatsApp in India) / Nudge message to reduce frictionless communication ("Read before sharing", on Twitter) / Counter-speech recommendation.

Platforms' response: // Mechanisms to take down content or reduce the dissemination / A special appeal process for journalists (as planned in the UK Online Safety Bill).

As a result of their work, Lorna Woods and William Perrin drafted a [Hate speech Code of practice with Guidelines and a series of recommendations on 12 areas](#), recommending flexible rules to be applied and adapted by platforms to their services, under the supervision of regulatory authorities.

How to proceed as a regulator?

- Understanding and measuring harms: It is crucial to **liaise with citizens and especially, victims and with a wide diversity of victims to avoid overlooking some distinctive experiences** - to learn from their experience and understand better what specifically affects them. Regulators should build on their experience of a citizen-centred regulation to achieve this.

Focus - the importance of a victim-focused approach:

When developing the hate speech code, the researchers spoke with a wide-range of charity organisations focused on helping specific categories of victims, (e.g. [CST](#)/anti-Jewish, [Glitch](#)/black women, [TellMama](#)/anti-muslim etc), they proved a key information source. Engaging with a variety of groups helps to understand better the nature of the problem and acceptable and proportionate solutions, especially with regard to dealing with non-criminal, slur-like, issues of hate speech (e.g. slowing the flow down etc.).

- An informed supervision: This approach has also a strong emphasis on the capacity for regulators **to hold and to request information**. Regulators shall have access to data, independent research resources and skilled staff to be able to challenge platforms and make them comply.

Challenges raised:

- The difficulty to reach some less visible categories of victims (if there are no (or a lack of) active groups, it suggested to run focus groups).

- The difficult for regulators to identify harms and people reactions before they occur: Industry has experience in the field of risk assessment and the idea is not to be accurate but to ensure that platforms have taken into account the likelihood of harm and the safety of users - *what is foreseeable based on the current evidence and available research* – when designing their services.

- The need of a shift of mindset to implement comprehensive safety tests in every provider's services.

However, a fast and evolving mindset is now becoming visible. For instance, before being implemented by major platforms, users' responses and tools to take back control on recommender algorithms were considered

by platforms in 2016 as against the spirit of social media. Moreover, anticipating harm remains less costly for platforms than to implementing remedies after the harm occurs.

- The issue of the political involvement and the attempt of governments to interfere in the regulator's organisation.
- The issue of the multinational context and the safeguard of human rights: Objective standards need to be applied by regulators but at some point, judicial bodies shall have the final word.
- The issue of harmful but non-criminal content: The legal provisions have to take into account other levels of hate speech content to efficiently tackle the impact of online hatred.

Lorna Woods suggests four categorisations of content: (1) content contrary to criminal law; (2) content contrary to regulatory provisions; (3) content that could give rise to a private cause of action; and (4) content that doesn't trigger any response under the law. **NB:** In Slovakia, with the new law and for regulatory purpose only, some legal components of a criminal offence are taken off to focus only on the content itself and its harmful consequences.

→ On the definition and categorisation of content, see also the [Council of Europe's Recommendations on combating hate speech](#).

To keep in mind:

- The workability and practicality of the systemic approach: The "why" and "how" might differ, but the fundamental basis of such a systemic approach seems to be the same and a universal code, no matter what the content domain is, could be possible (see the [Violence Against Women and Girls Code of Practice](#)).
- A regular assessment and monitoring: New challenges arise all the time and the systems shall be regularly assessed and improved if necessary (repeat cycle: responsibility, risk assessment, mitigation and remediation).
- The proportionality of remedies and answers to hate speech: Some harmful content are criminal offences and require effective responses to serious threats, while some of the "lower" offences could be efficiently addressed with safer or more incentive communication processes.

→ **For more details, see the presentation of Lorna Woods and William Perrin:**

<https://www.epra.org/attachments/vsps-regulation-workshop-n-3-presentation-by-lorna-woods-and-william-perrin>

○ **Brief overview of relevant experiences from EPRA members and observers:**

➤ **Regulators' initiatives in non-EU countries:**

- Ofcom, UK by Murtaza Shaikh: in October 2021 Ofcom has published its final [guidance for video-sharing platform providers on measures to protect users from harmful material](#). The UK regulator is now in the phase of engaging with platforms, with a priority focus on hatred and terrorism on relevant major platforms. This guidance is a pilot test for the Online Safety Bill and Ofcom will publish its first **public annual report** on video-sharing platforms in Autumn 2022. In the meantime, Ofcom will release its **roadmap to regulation**, setting out Ofcom's regulatory approach for

implementing online safety regulation and including a sector-wide overview of the harms identified in the platform industry.

Update 06/07/2022: Publication of the [roadmap](#) together with a [call for evidence](#).

- CRA, BA by *Maida Culahovic*: The AVMSD transposition is still pending but online media and especially hate speech are the most prominent issue in the media environment. A complex political situation and crisis make any legislation process difficult, but the regulator is trying to be proactive. The CRA has already initiated some coalition and a cooperation platform with the stakeholders and keeps advocating for such cooperation. So far, positive feedback was received and a study due in October within the [JUFREX](#) programme should map the relevant players (*institutions, regulators...*), assess their capacities to be involved in joint-responsibilities and provide recommendations for the establishment and functioning of such a collaborative platform. This work could be used for the Digital Services Act's implementation at a later stage. Some early steps are also made through a coordinated regulatory approach at regional level with the non-EU neighbouring countries to give the regulators a stronger voice.

➤ **New judicial and legislative developments in EU countries:**

- DLM, DE by *Peter Matzneller*: In Germany, the [Network Enforcement Act](#) (NetzDG) aimed at tackling online hate speech, has entered into force in 2018. This law obliges major social media platforms to remove manifestly illegal content within 24 hours and illegal content within 7 days after being reported, and to provide regular reporting. However, in the case of a lawsuit filed by Alphabet and Meta, on 1 March, the [Administrative Court of Cologne](#) held that recent reporting obligations added to the NetzDG were partially not in line with European law and especially with the country-of-origin principle and, consequently, the parts in violation may not be applied by the national authorities until a final judgment in the main proceedings. Moreover, the court also ruled that the competent authority appointed by law (Federal Office of Justice) is not fully independent, as required by European law. The German government will now have to decide to amend, if possible, or to abolish the law.
- KommAustria, AT, by *Daniel Schärf*: In Austria, the enforcement of the [Communication Platforms Act](#) (Kommunikationsplattformen-Gesetz - KoPI-G) raised similar issues than in Germany (compliance with the EU country-of-origin principle). On 24 May 2022, the Supreme Administrative Court referred a [preliminary ruling](#) to the European Court of the European Union to decide on the compliance of the KoPI-G with European law, thus suspending current decisions against Meta and Alphabet. In general, platforms comply with most of the provisions while legally fighting every provision they can. However, the law has allowed for the first time to collect data and to get an overview of the type of content and offences reported.
- CBR, SK, by *Stanislav Matějka*: New legislation transposing the [AVMSD](#) has just been adopted in Slovakia and will enter into force on 1st August. With the [new law](#) the CBR becomes the Council for Media Services and it grants the regulator with wider competence for the oversight of the measures undertaken by VSPs and to monitor online criminal hateful content. Platforms will have the obligation to implement complaint mechanisms, and, in case of failure, users will have the possibility to submit a claim to the regulator. The problems notified to the media authority will be publicly published on the authority's website. The CBR has already built relationships and exchanges with

platforms and relies on its experience and network (*academics, citizens associations...*) to be able to effectively analyse and assess the platforms' data collected.

➤ **Open consultation on the access to platform's data:**

- Arcom, FR by *Sébastien Lécou*: Arcom has launched an online open consultation on the collection of platforms' data and the issues encountered by researchers in this regard. To understand the role played by online platforms in the spread and effects of hate speech, research is key. The consultation, not specifically focused on hate speech but on the access to platforms' data in general, aims at understanding how researchers conduct studies, collect data and what kind of difficulties they encounter. The consultation is available in French and English.

→ Link to the consultation: <https://www.arcom.fr/consultations-publiques/consultation-publique-sur-lacces-aux-donnees-des-plateformes-en-ligne-pour-la-recherche>

➤ **Council of Europe's Recommendation on combating hate speech:**

- Council of Europe by *Urška Umek*: The Committee of Ministers of the Council of Europe has adopted last May a recommendation to encourage changes to hate speech legislation. The recommendation provides a comprehensive approach to hate speech, defined through categories based on the severity of their impact. The recommendation acknowledges that elements of hatred can be found in various content which may not automatically represent criminal offences but may fall under civil or administrative provisions. They refer to a number of non-legal measures (*education, media literacy, counter speech...*) to respond to harmful content depending on their severity.

➤ **A selection of relevant reading material**

On the systemic approach to tackle hate speech on online social media (Carnegie Project):

- Ad hoc advice from Carnegie UK to United Nations Special Rapporteur on Minority Issues, *Professor Lorna Woods, William Perrin, Maeve Walsh, Carnegie UK, 2022*:
https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2021/07/25105219/UN-Hate-Speech-draft-v.05a-1.pdf
- Violence Against Women and Girls (VAWG) Code of Practice, *The End Violence Against Women Coalition, Glitch, Refuge, Carnegie UK, NSPCC, 5Rights, Professor Clare McGlynn and Professor Lorna Woods, Carnegie UK, 2022*:
https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2022/05/24163713/VAWG-Code-of-Practice-16.05.22-Final-1.pdf

See also:

- Online harm reduction – a statutory duty of care and regulator, *Professor Lorna Woods & William Perrin, Carnegie UK, April 2019*:
https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2019/04/06084627/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf

- The Carnegie Project "*Tackling Online Harms*":
<https://www.carnegieuktrust.org.uk/programmes/tackling-online-harm/>

On the issue of access to data and transparency:

- Report of the European Digital Media Observatory's (EDMO) Working Group on Platform-to-Researcher Data Access, including a draft Code of Conduct on how platforms can share data with independent researchers while protecting users' rights, *May 2022*:
<https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf>
- ARCOM's (FR) public consultation on access to data from online platforms for research purposes; *May 2022*:
<https://www.arcom.fr/consultations-publiques/consultation-publique-sur-lacces-aux-donnees-des-plateformes-en-ligne-pour-la-recherche> // [English version]

On the international principles applying to risky industries:

- The UN Guiding Principles on Business and Human Rights:
<https://www.unglobalcompact.org/library/2>
- OECD Due Diligence Guidance for Responsible Business Conduct:
<https://www.oecd.org/investment/due-diligence-guidance-for-responsible-business-conduct.htm>

➤ Agenda



VSPs & Regulation Workshop – Session Nr. 3

A systemic approach to hate speech and the challenges of practical application

Agenda






29 June 2022



Virtual coffee chat opens: 9:45 CET



Meeting starts: 10:00 CET

10:00 – (10 min)	Welcome & introduction - <i>Luboš Kukliš, EPRA Chairman</i>
	 Presentation of the topic, the aim & structure of the 3 rd workshop
10:10 – (50 min)	Part 1 – Academic inputs and discussion Benefits of a systemic approach to hate speech
	 Prof. Lorna Woods , University of Essex (10-15 min)  William Perrin , Carnegie UK Trust (10-15 min)  Q&A with the floor (20 min)
11:00 – (50 min)	Part 2 – Regulators' roundtable Exploring the challenges of practical regulatory application
	 Open discussion based on short interventions covering issues such as: <ul style="list-style-type: none"> • <i>How would such a systemic approach work in practice</i> • <i>The role that media NRAs can play in this collaborative governance</i> • <i>First experiences of media NRAs with liaising with VSPs on hate speech issues</i> • <i>First experiences of media NRAs with developing guidance on hate speech</i> • <i>Resources (or the lack of) that VSPs put into safety & content moderation</i> • <i>Primacy of the English language in safety & content moderation</i> • <i>Understanding the systems: the issue of transparency & access to platforms' data</i>
11:50 – (10 min)	Closing remarks and next steps - <i>Luboš Kukliš, EPRA Chairman</i>



Meeting closes: 12:00 CET