Potential impact of the DSA on disinformation

REMEDIS - REgulatory and other solutions to MitigatE online DISinformation

Uni.Lu

Huiyun Tang
Yuwei Chuai
Anastasia Sergeeva
Gabriele Lenzini

UCLouvain

Jean De Meyere
Luc Desaunettes-Barbero
Alain Strowel
Daphné Chapellier
Marie Dufrasne

UNIVERSITÉ DU LUXEMBOURG
UCLouvain
UCLouvain SAINT-LOUIS BRUXELLES
fnrs LA LIBERTÉ DE CHERCHER
Luxembourg National Research Fund
Remedis
REgulatory and other solutions to Mitigate online DISinformation

The Remedis Team

Multidisciplinary Experts

Specialists from digital law, Computer Science,Human-Computer Interaction, Communication Science

Our Mission

Combating online dis/misinformation with innovative solutions, ensuring a balance between freedom of expression and social responsibility in the digital world.



UNIVERSITÉ DU LUXEMBOURG

UCLouvain

UCLouvain SAINT-LOUIS BRUXELLES

fnrs LA LIBERTÉ DE CHERCHER

Luxembourg National Research Fund

Agenda

1. Risk assessment obligation (Art. 34 DSA) and detection of disinformation from a technical and cybersecurity perspective
2. Curation of disinformation under the DSA from a legal perspective
3. Mitigation measures (Art. 35 DSA) from the perspective of Human-computer interactions science
4. Extra-legal remedies from the perspective of communication sciences

The DSA

- **Adopted the 19 October 2022**
- **Tiered structure : more obligations for large platforms (VLOPs: >45M EU users)**

Objectives

- Protecting rights of users online
- Ensuring a safe, predictable and trusted online environment
- Provide the appropriate supervision of digital services

No focus on disinformation but

- Strenghtened transparency obligations for online platforms
- VLOPs need to analyse and address systemic risks, including those favoring disinformation

Risk assessment obligation (Art. 34 DSA) and detection of disinformation from a technical and cybersecurity perspective

Article 34

Article 34 Risk assessment

1. Providers of very large online platforms and of very large online search engines shall diligently identify, **analyse and assess any systemic risks in the Union stemming from the design or functioning of their service** and its related systems (...)

This risk assessment (...) shall include the following systemic risks: (...)

(b) any actual or foreseeable negative effects for the **exercise of fundamental rights,** in particular the fundamental rights to human dignity enshrined in Article 1 of the Charter, (...), to freedom of expression and information, including the freedom and pluralism of the media, enshrined in Article 11 of the Charter (...).

(c) any actual or foreseeable negative effects on civic **discourse and electoral processes**, and public security;

(d) any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being

Article 34

2. When conducting risk assessments, providers of very large online platforms and of very large online search engines shall take into account, in particular, **whether and how the following factors influence any of the systemic risks** referred to in paragraph 1:

(a) The design of their recommender systems and any other relevant algorithmic system;
(b) Their content moderation systems;
(c) The applicable terms and conditions and their enforcement;

The assessments shall also analyse whether and how the risks pursuant to paragraph 1 are influenced by **intentional manipulation of their service,** including by inauthentic use or automated exploitation of the service, **as well as the amplification and potentially rapid and wide dissemination** of illegal content and of **information that is incompatible with their terms and conditions.**

Machine learning based methods

Content

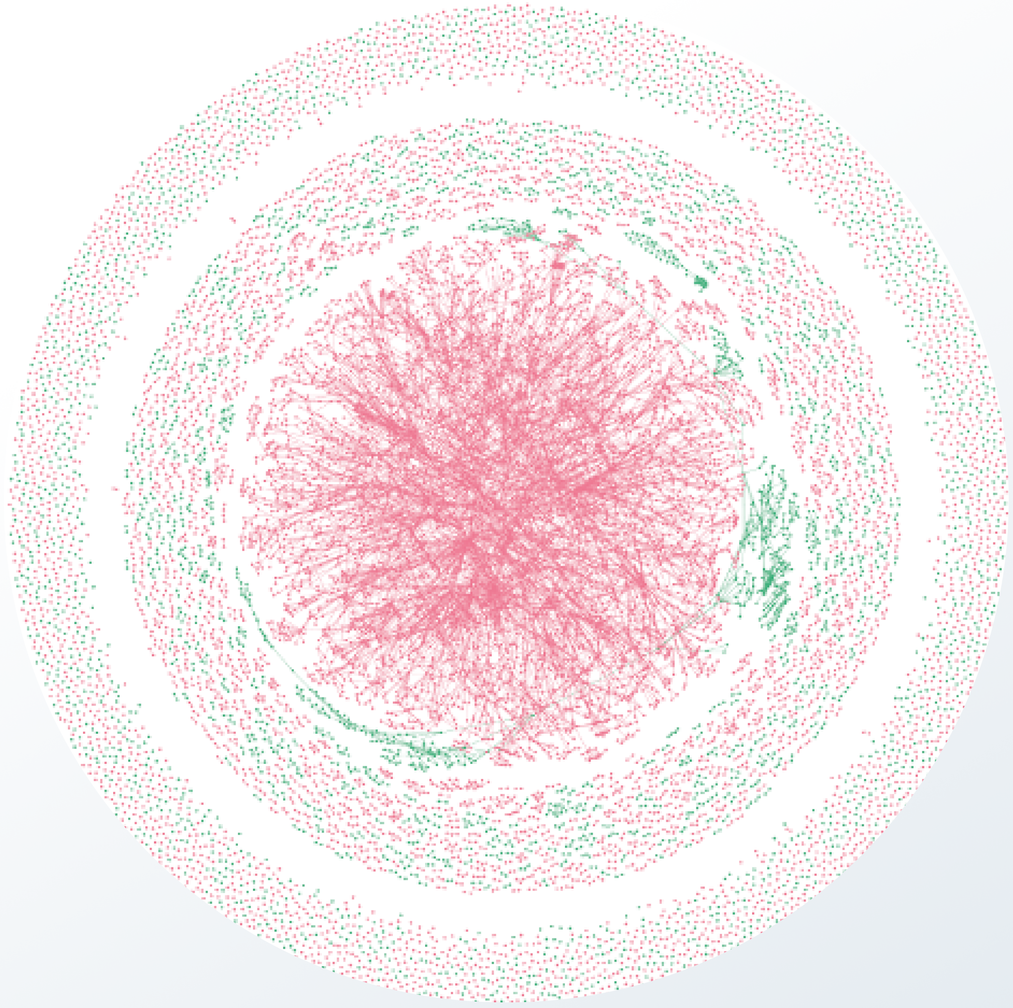
- **Knowledge-based:** External knowledge (Fung et al., 2021)
- **Style-based:** Emotions (Shu et al., 2019), Topics (Hu et al., 2022), etc
- **Multi-modal based:** Combining text and image (or video) features as input (Singhal et al., 2022)

Context

- **Aggregated-based:** Incorporation content and context features together as input (Das et al., 2022)
- **Network-based:** Diffusion network (Wei et al., 2022), Cooccurrence network (Dou et al., 2022), Hierarchical network (Jin et al., 2021)

Lack of counter-evidence in automatic detection

word co-occurrence

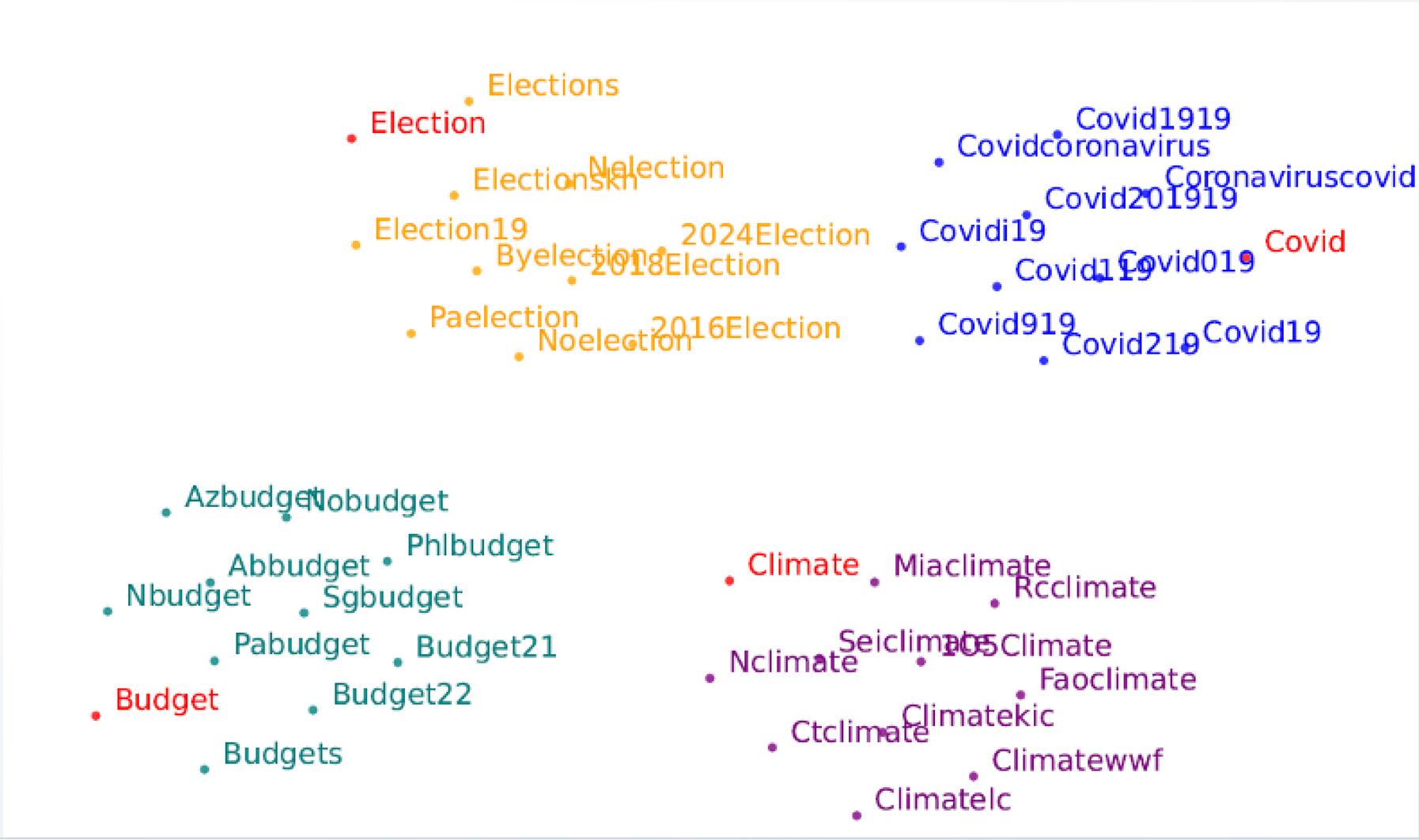


True

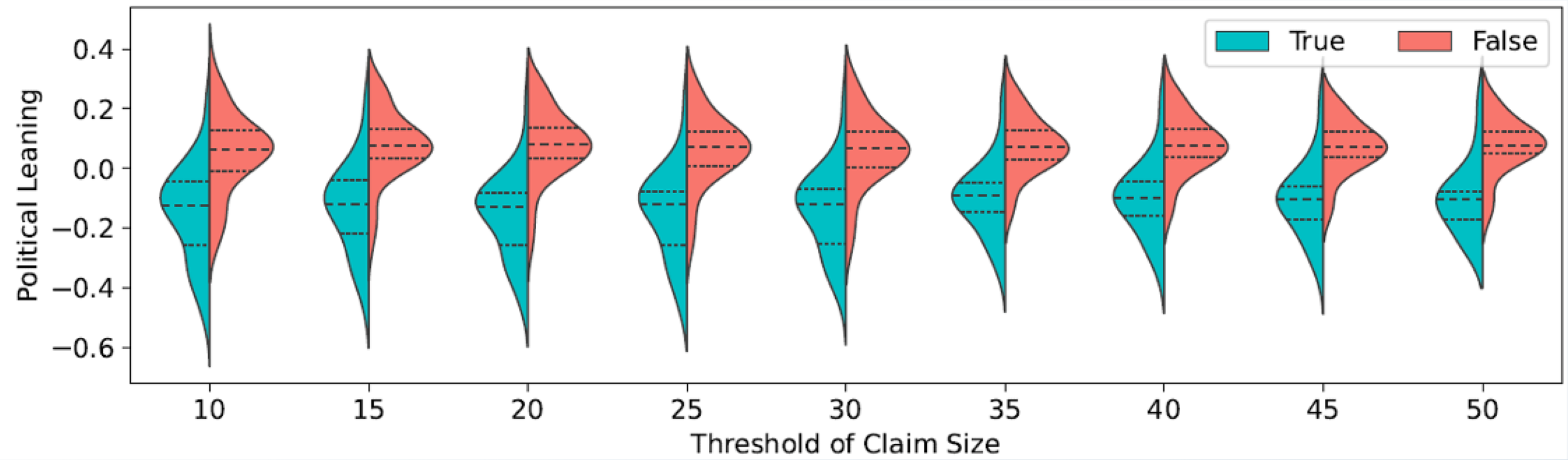
False

Word Co-occurrence network without common words.

Word Clustering



Political leanings



In practice

Tiktok

Redirection strategy

Youtube/Google

Knowledge panel

Meta

Third-party fact-checker

Community notes



Prebunking

- Security warnings
- General links

Need more targeted and tailored responses

Debunking

- Fact-checking by experts
- Automatic Detection

Accuracy and large-scale

The Roll-Out of Community Notes Did Not Reduce Engagement With Misinformation on Twitter

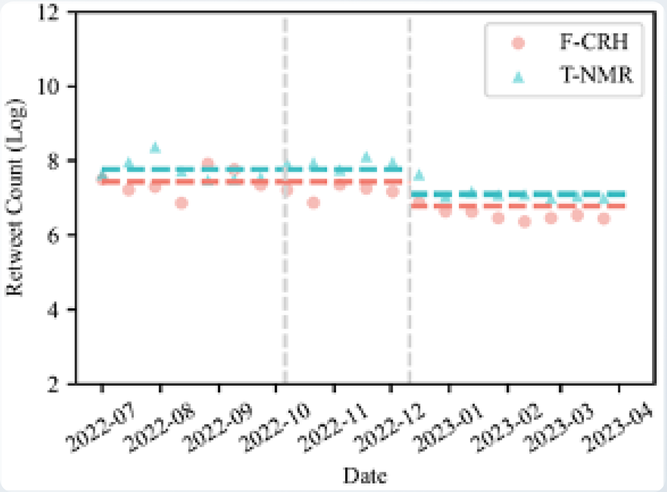
Chuai, Y., Tian, H., Pröllochs, N. and Lenzini, G.

We perform a large-scale empirical study to analyze the effectiveness of Community Notes feature.

(i)We find no evidence that the introduction of **Community Notes significantly reduced engagement** with misleading tweets on Twitter.

(ii)Rather, our findings suggest that Community Notes might be **too slow to effectively reduce engagement** with misinformation in the **early** (and most viral) stage of diffusion.

Our work emphasizes the importance of evaluating fact-checking interventions in the field and offers important implications to enhance crowdsourced fact-checking strategies on social media.



scan to read the full paper

Curation of disinformation under the DSA from a legal perspective

Disinformation in the DSA?

No definition of disinformation in the DSA!

Definition of illegal content – DSA art. 3 : any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law;

No definition and prohibition of “disinformation” at EU level but certain cases of disinformation might fall under prohibited offenses (ex: harassment, hate speech, etc.)

→ Disinformation = “lawful yet awful” content

Regulation focuses on the safeguards platforms should put in place regarding public safety and fundamental rights

Regulating how platforms regulate speech

Article 14 – Terms and conditions

1. Providers of intermediary services shall include information on **any restrictions** that they impose in relation to the use of their service in respect of information provided by the recipients of the service, in their terms and conditions. That information shall include information on **any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review, as well as the rules of procedure of their internal complaint handling system.** It shall be set out in clear, plain, intelligible, user-friendly and unambiguous language, and shall be publicly available in an easily accessible and machine-readable format (...)

4. Providers of intermediary services **shall act in a diligent, objective and proportionate manner in applying and enforcing the restrictions** referred to in paragraph 1, with **due regard to the rights and legitimate interests of all parties involved**, including the fundamental rights of the recipients of the service, such as **the freedom of expression, freedom** and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter.

Normative power in the DSA

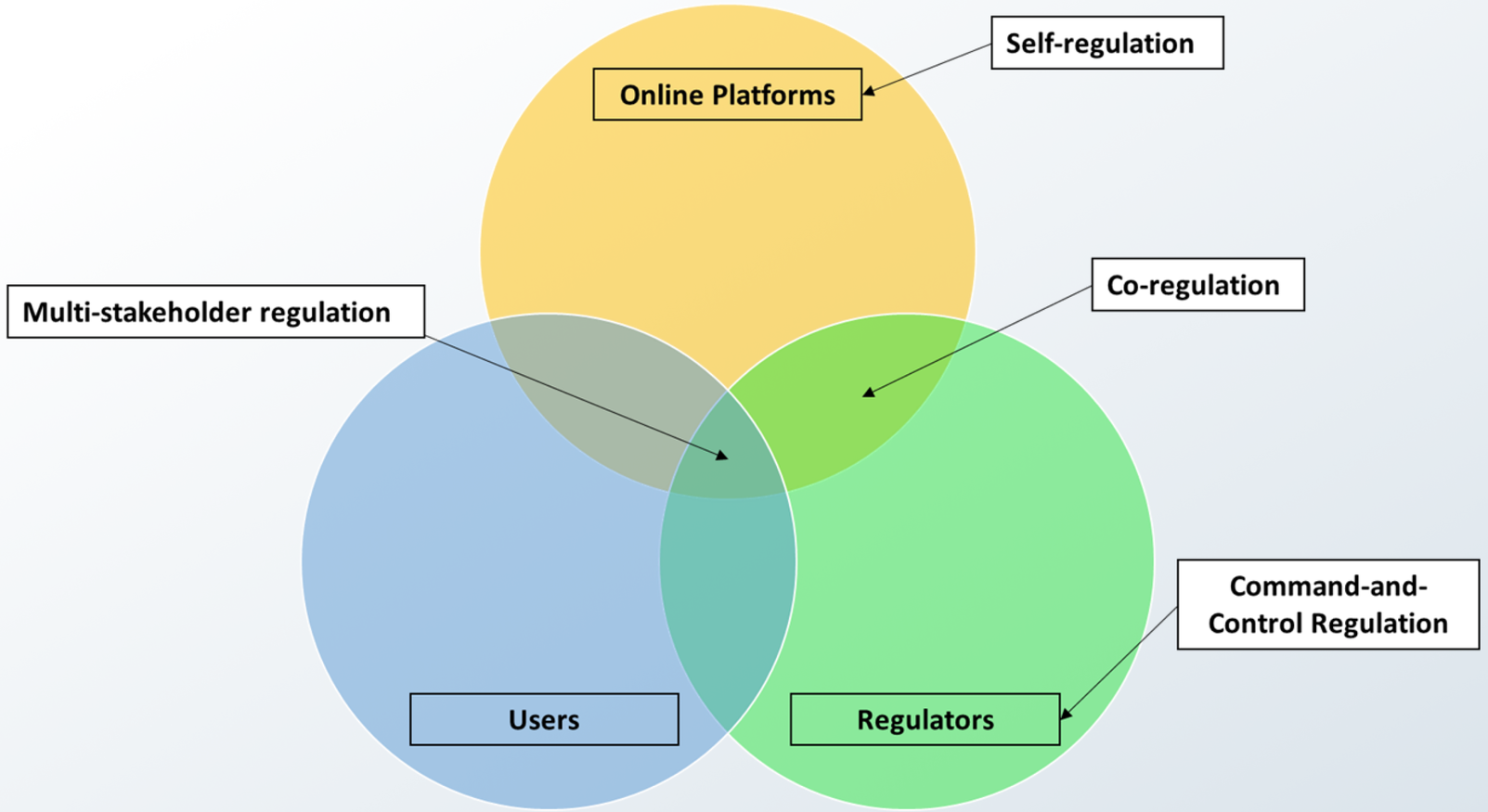
Platforms are still responsible for determining which content they moderate and how !

→ Under the supervision of the EU Commission and DSC

Potential improvement regarding transparency & accountability… however :

- Legitimacy question still open
- EC tells platform what to do, not how to do it
- Normative power remains largely in the hands of the platforms – what about users?

Towards multi-stakeholder regulation?



Mitigation measures (Art. 35 DSA) from the perspective of Human-computer interactions science

DSA – article 35 – mitigation measures

Article 35 Mitigation of risks

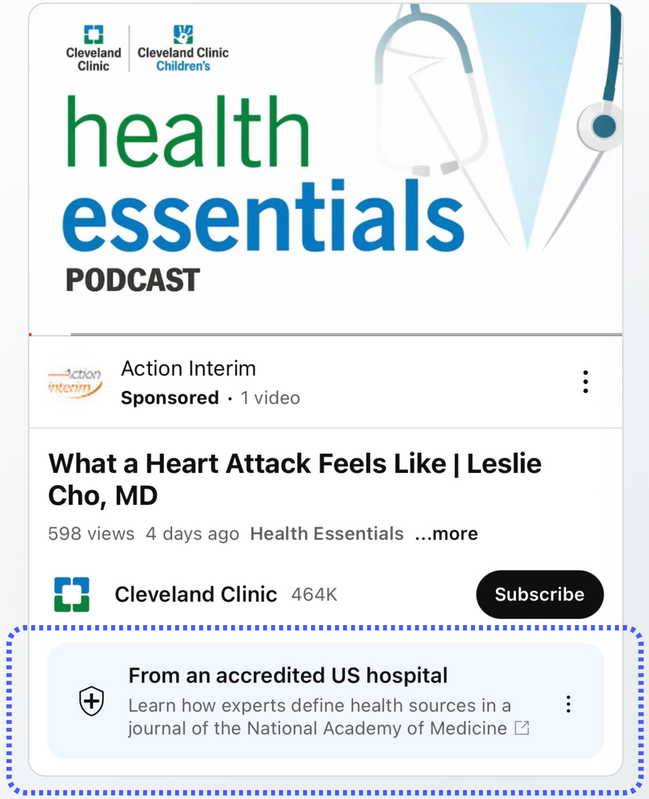
1. Providers of very large online platforms and of very large online search engines **shall put in place reasonable, proportionate and effective mitigation measures**, tailored to the specific systemic risks identified pursuant to Article 34, **with particular consideration to the impacts of such measures on fundamental rights.**

Such measures may include, where applicable:
(a)**adapting the design, features or functioning of their services, including their online interfaces;**
(b)adapting their terms and conditions and their enforcement;
(c)adapting content moderation processes, including (…), where appropriate, the expeditious removal of, or the disabling of access to, the content notified, in particular in respect of illegal hate speech or cyber violence, as well as adapting any relevant decision- making processes and dedicated resources for content moderation;
(d)testing and adapting their algorithmic systems, including their recommender systems;
(…)
(k) ensuring that an item of information, whether it constitutes a generated or manipulated image, audio or video (…) is distinguishable through prominent markings when presented on their online interfaces, and, in addition, providing an easy to use functionality which enables recipients of the service to indicate such information.

Current anti-Dis/misinformation solutions

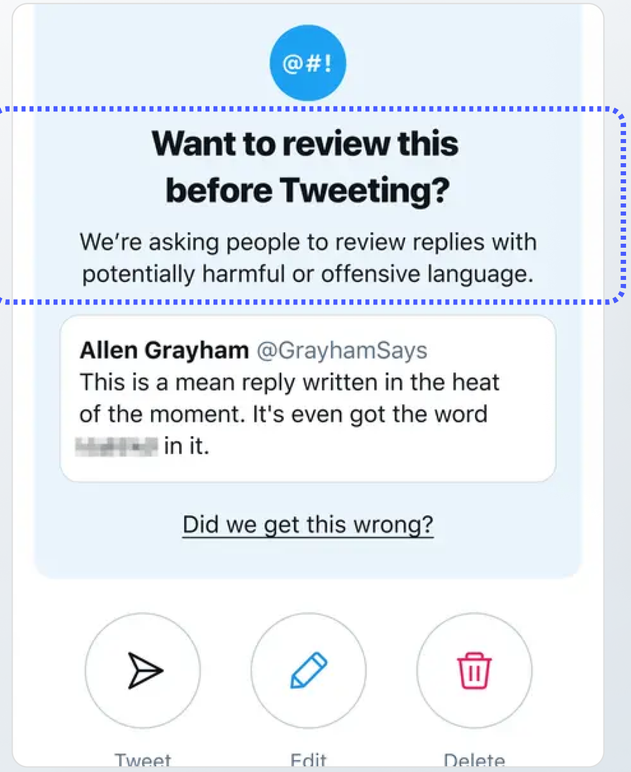
Credibility indictors

assist users in assessing the credibility and reliability of the media they consume



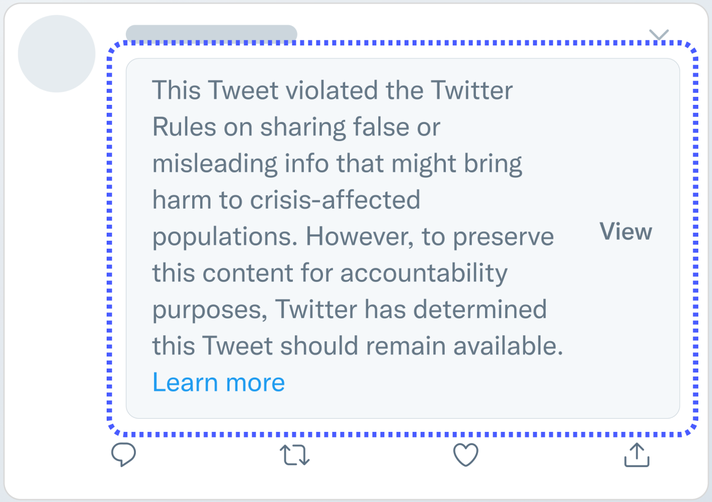
Nudges

raise awareness and promote critical news consumption behavior



Reducing Content Visibility

Reduce the visibility of misleading content



Understanding User Perceptions of Misinformation Countermeasures

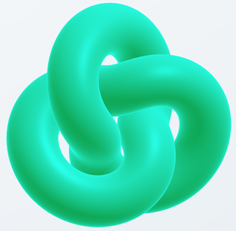
we conducted 22 in-depth interviews, focusing on perceptions and platform responsibilities.



Oversimplification of message credibility



Lack of **transparency**



Lack of **user agency**



Did not consider the **features** of different kinds of (mis)information

The Impact of Generative AI on Dis/Misinformation Spread

We conducted an extensive study with 300 participants to explore how AI-generated misinformation affects public opinion and the efficacy of fact-checking measures.

- People tend to **trust the algorithmic systems**, even it's **obviously wrong**
- Current countermeasures against AI-generated misinformation are **insufficient**.
- It is imperative to increase **transparency** and **understanding** of AI systems.

scan to read the full paper

Extra-legal remedies from the perspective of communication sciences

Extra-legal remedies

Print press organizations and broadcasters are intensifying their efforts to implement **trust-enhancing practices**, such as fact-checking (Shattock, 2021).

Limited impact explained by

- the potential "boomerang effect" (backlash) correcting methods tend to cause (Scheufele & Krause, 2019).
- the exposure to fact-checked articles is limited (Altay, 2022).



Extra-legal remedies

Media and information literacy (MIL):

- Efforts should be made to promote media and information literacy practices to educate the general public, **enhancing resilience among end-users and citizens in all age groups** (Shattock, 2021).
- MIL enables the teaching of fundamental skills in a digital age for **fostering critical thinking** and **responsible online behaviour.**

Media literacy is a **complementary solution** that goes hand in hand with regulatory measures and countermeasures.

Conclusion

Remedis

REgulatory and other solutions to Mitigate online DISinformation

Potential impact of the DSA on disinformation

REMEDIS - REgulatory and other solutions to MitigatE online DISinformation

Thank you!

remedis.project@uni.lu

Remedis Project

UNIVERSITÉ DU LUXEMBOURG

UCLouvain

UCLouvain SAINT-LOUIS BRUXELLES

fnrs LA LIBERTÉ DE CHERCHER

Luxembourg National Research Fund

Remedis
REgulatory and other solutions to Mitigate online DISinformation