

Online Filtering

Andrea Loreggia

Giovanni Sartor

Introduction

- Digital Services Act (DSA)
 - regulation of digital services
 - online platforms
- User-generated content:
 - enable users to express themselves
 - create, transmit or access information and cultural creations
 - engage in social interactions.



What is moderation?

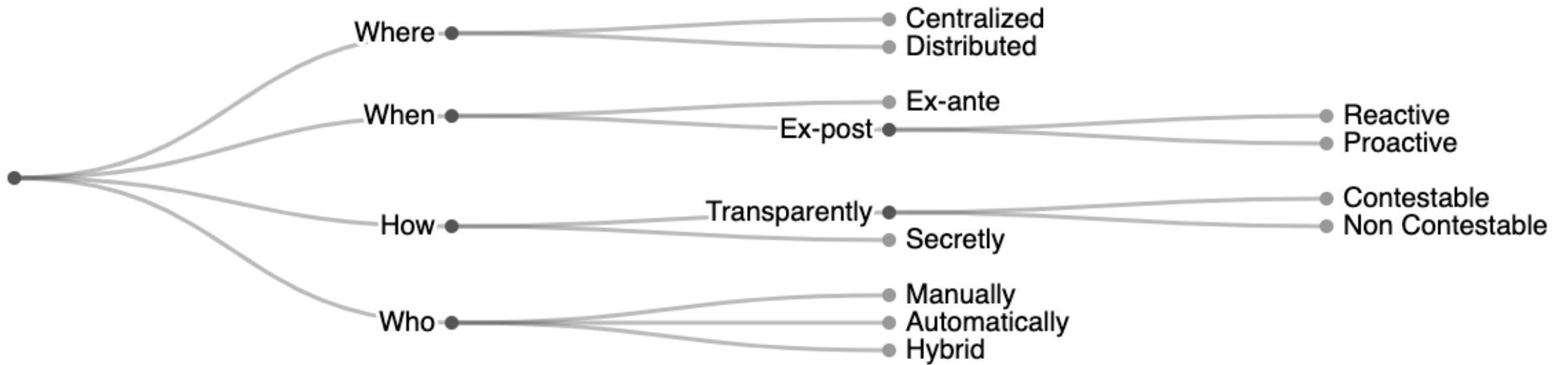
- Moderation is the active governance of platforms meant to ensure interactions among the users that are:
 - Productive
 - Pro-social
 - Lawful



Why filtering?

- To prevent unlawful and harmful online behaviour
- To mitigate its effect
- To facilitates cooperation
- To prevents abuse

Taxonomy



Taxonomy - Where

- *Centralized filtering*, which is applied by a central authority according to uniform policies, that apply to a whole platform.
- *Decentralized filtering*, which involves multiple distributed moderators, operating with a degree of independence, and possibly enforcing different policies on subsets of the platform.

Taxonomy - When

- *Ex-ante filtering*, which is applied before the content is made available on the platform.
- *Ex-post filtering*, which is applied to the content that is already accessible to the platform's users.

Taxonomy - When

- *Ex-ante filtering*, which is applied before the content is made available on the platform.
- *Ex-post filtering*, which is applied to the content that is already accessible to the platform's users
 - *Reactive filtering*, which takes place after the issue with an item has been signalled by users or third parties.
 - *Proactive filtering*, which takes place upon initiative of the moderation system, which therefore has the task of identifying

Taxonomy - How

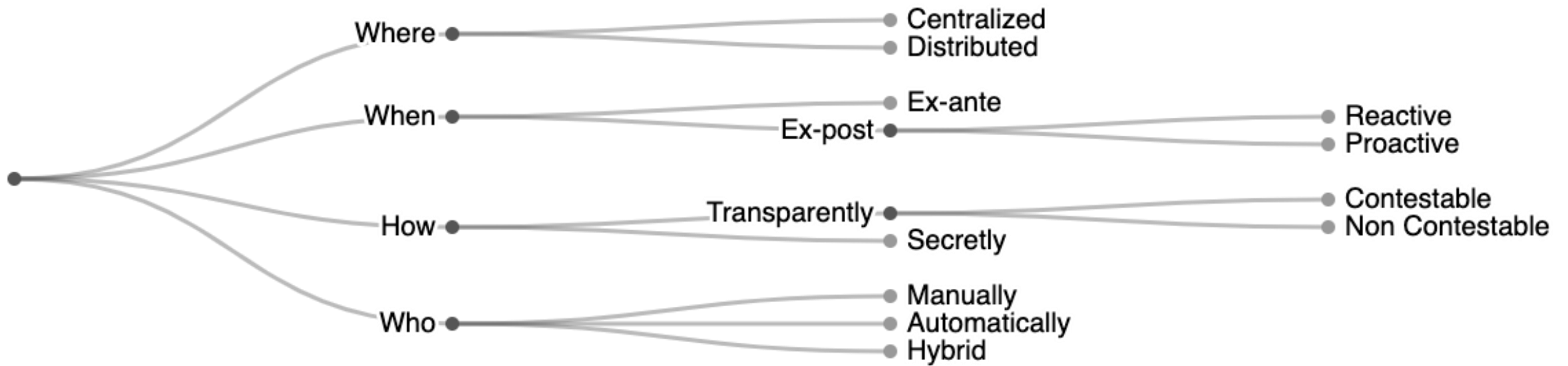
- *Transparent filtering*, which provides information on the exclusion of items from the platform.
- *Secret filtering*, which does not provide any information about the operation.

Taxonomy - How

- *Transparent filtering*, which provides information on the exclusion of items from the platform.
 - *Contestable filtering*. The platform provides uploaders with ways to contest the outcome of the filtering, and to obtain a new decision on the matter.
 - *Non-contestable filtering*. No remedy is available to the uploaders.
- *Secret filtering*, which does not provide any information about the operation.

Taxonomy - Who

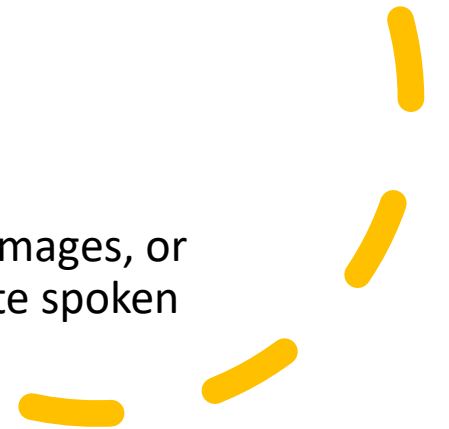
- *Manual filtering*, which is performed by humans.
- *Automated filtering*, which is performed by algorithmic tools.
- *Hybrid filtering*, which is performed by a combination of humans and automated tools.



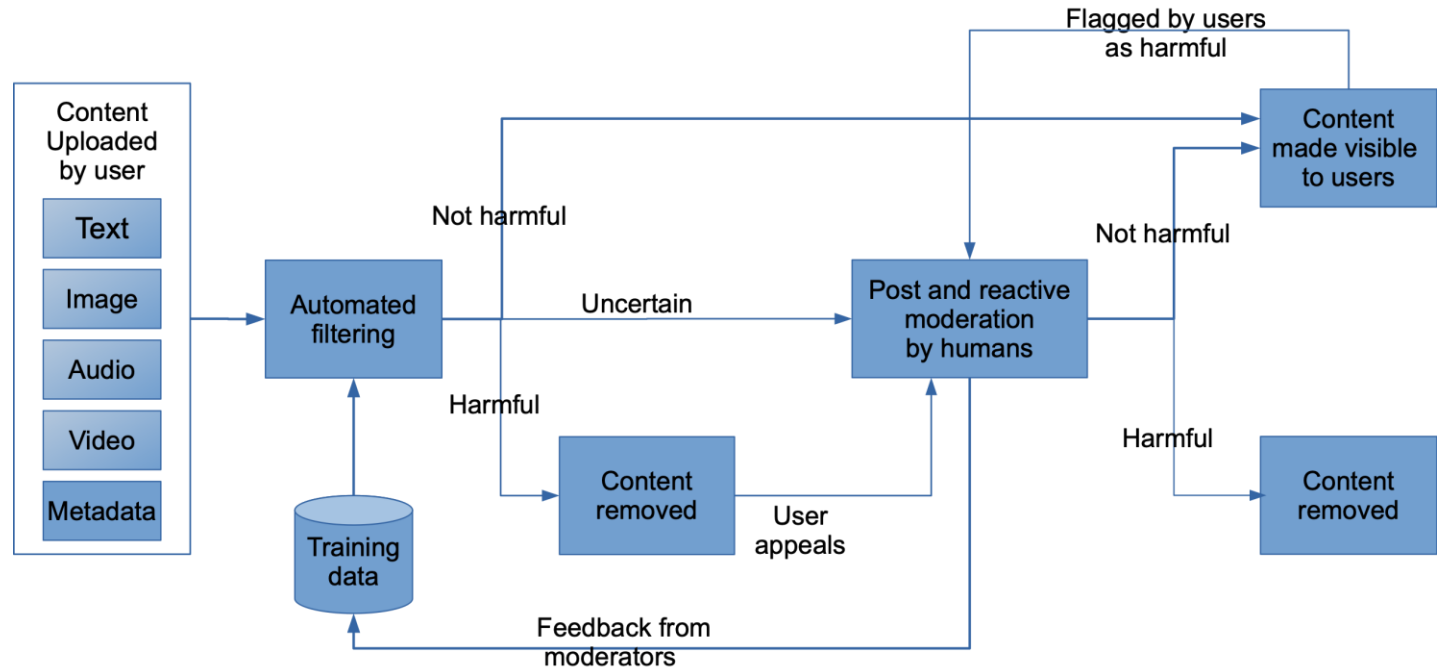


Different Media

- Metadata searching, hashing, and fingerprinting → to identify copies of known digital works;
- Blacklisting → to find unwanted expressions;
- NLP → to address meaning and context;
- Multiple AI techniques → to identify unwanted images, or combinations of text and images, and to translate spoken language into text.




How It Works



Some Examples

INDEPENDENT

Subscribe LOGIN



The Little Mermaid statue is one of Denmark's best-loved sights (ODD ANDERSEN/AFP/Getty Images)

FACEBOOK REMOVES IMAGE OF COPENHAGEN'S LITTLE MERMAID STATUE FOR BREAKING NUDITY RULES

CNN travel VIDEO Q



Curiosità e scorci di Bologna

Enza Barbara FANPAGE

Facebook banned Neptune statue photo for being 'explicitly sexual'

Sara Delgrossi and Lauren Said-Moorhouse, CNN • Updated 5th January 2017

Some Examples

ISSIE LAPOWSKY BUSINESS 03.15.2019 01:50 PM

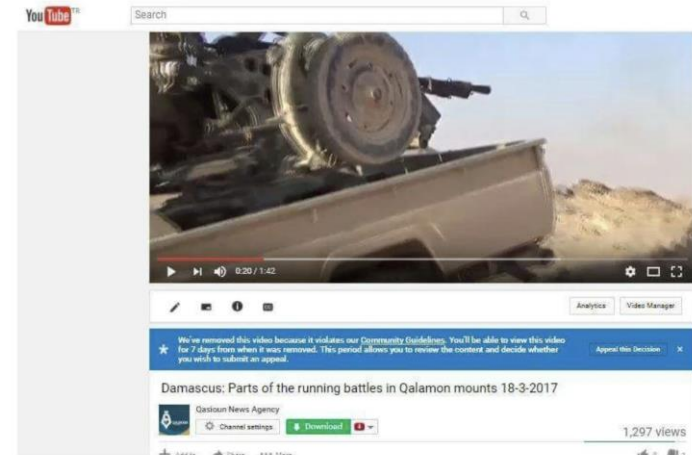
Why Tech Didn't Stop the New Zealand Attack From Going Viral

Video from mosque shootings in Christchurch popped up on Facebook, Reddit, Twitter, and YouTube, showing the limits of social media moderation.

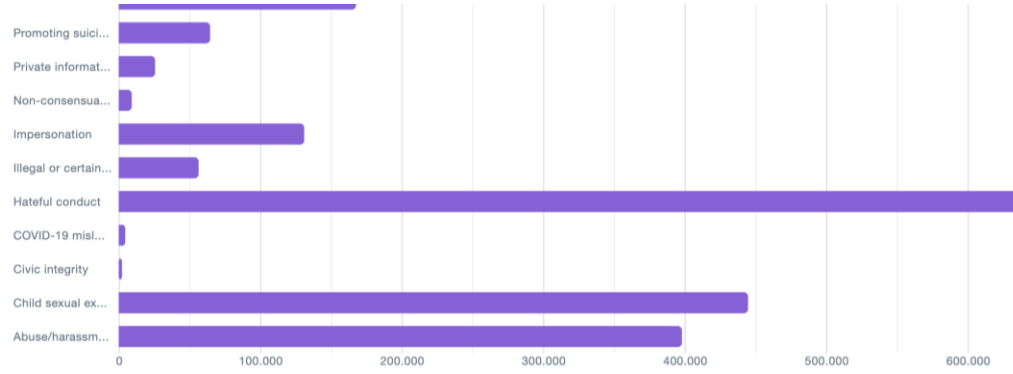


Creazione di una connessione protetta in...

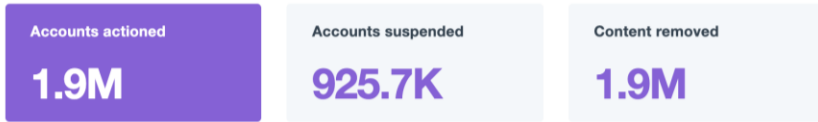
YouTube Removes Videos Showing Atrocities in Syria



A takedown notice issued by YouTube on a video of the Syrian conflict. YouTube



Accounts actioned - January - June 2020



Transparency

Example from Twitter transparency report



Not of filtering aim
at moderation

- Filter bubbles
- Echo chambers
- Censorship
- Fake news

Filtering and the law

- Does/Should the law allow filtering?
 - Is/Should there be liability for automatically filtering out/removing content?
- Does/Should the law impose filtering?
 - Is/Should there be liability for for not having automatically filtered out content
- What about ethics?
 - Does ethics require us to filter out/remove content?
 - Does ethics prohibits us to filter out/remove content?

The technological/legal challenge

- Can a provider be obliged to what is impossible? Ad impossibilia ...
But technology can make the impossible possible.
- For a legal obligation to filter out content to exist, it must be the case that technologies are available which provide a positive trade-off between
 - The benefit of filtering out/removing unlawful/harmful content
 - The drawback of filtering out/removing lawful/beneficial content

What does it mean to have control

- When we say that providers have control over the content on their premises it means that
 - They have the ability to intervene filtering out, or remove, given the technologies that they have adopted (real control), without causing unacceptable side effects
 - They would have had that ability if they had adopted available/accessible technologies (potential control).
- Failure to exercise control (including failure to acquire that ability), may lead to liability, relatively to unlawful content

The current regulatory framework: eCommerce directive

- Article 14. host providers are exempted from liability as long as they
 - do “not have actual knowledge of illegal activity or information” and, as regards claims for damages,
 - are “not aware of facts or circumstances from which the illegal activity is apparent
 - upon gaining awareness act “expeditiously to remove or to disable access to the information
- Article 15 of the eCommerce directive prohibits Member States from imposing any
 - “general obligation on providers [...] to monitor the information which they transmit or store” or any general obligation to “actively to seek facts or circumstances indicating illegal activity
 - But what obligation is general?

What objects for an obligation to filter out/remove?

- a single copy of unlawful file (e.g., a downloadable video or a web page) being identified by an univocal identifier (e.g., its URL, the universal locator for resources available online);
- all copies of an individually identified file that are present in on an online repository (e.g., all copies of a certain copyrighted video);
- not only the copies of a file that currently made accessible by the provider, but all subsequent reposting of further copies;
- not only the copies of a given unlawful file, but also all variations obtained from that file, while keeping its unlawful meaning;
- all content that is relevantly similar to, and equally unlawful as, a specified unlawful file, even when such content is not obtained by changing that document (e.g., all messages conveying a certain kind of defamatory or hateful content relatively to certain individuals,);
- all documents that are unlawful in virtue of certain common features (e.g., unauthorised reproduction of copyrighted works contained in a given database, paedophilia, incitation to hatred, or terrorism, etc.)

What is general

- "generality" of a measure refers to the imbalance between the advantages and disadvantages of implementing a removal/filtering obligation with regard to
 - the extent to which the measure would effectively contribute to prevent unlawful content and the gravity of the harm that distributing such content might cause;
 - the extent to which the measure would also filter out permissible content and the gravity of the harm that non distributing such content might cause;
 - the sustainability of the costs that the measure may impose on providers;
 - the extent to which it may affect the usability of a platform or user engagement in a community.

Obligation to filter/remove

- Audiovisual Media Services Directive, at Article 28b (1),
 - providers should take appropriate measures to protect minors from content that may impair their development, and the general public from content that incites to violence or hatred, or whose dissemination is a criminal offence in connection with terrorism, child pornography, racism and xenophobia.
- Copyright directive, at Article 17 (4)
 - online content-sharing service providers shall be liable for unauthorised acts of communication to the public, unless they providers demonstrate that they have reacted to notices removing the material and make best efforts to prevent future uploads

The digital services act: a new complex framework

- Obligation to implement orders to take action and provide information
- Point of contact for users' complaints, notice and action
- Statement of reason when removing or blocking
- Good Samaritan provision: liability exemptions consistent with activities aimed at identifying and removing, or disabling of access to, illegal content
- Online platforms (not small enterprises)
 - Complaint handling system, Transparency
- Very large online platforms
 - Assessment of risk concerning the dissemination of illegal content
 - Obligation to adopt reasonable, proportionate and effective mitigation measures

Thanks for your attention!

giovanni.sartor@unibo.it