# Ethics of Filtering

Andrea Loreggia

MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

# 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.

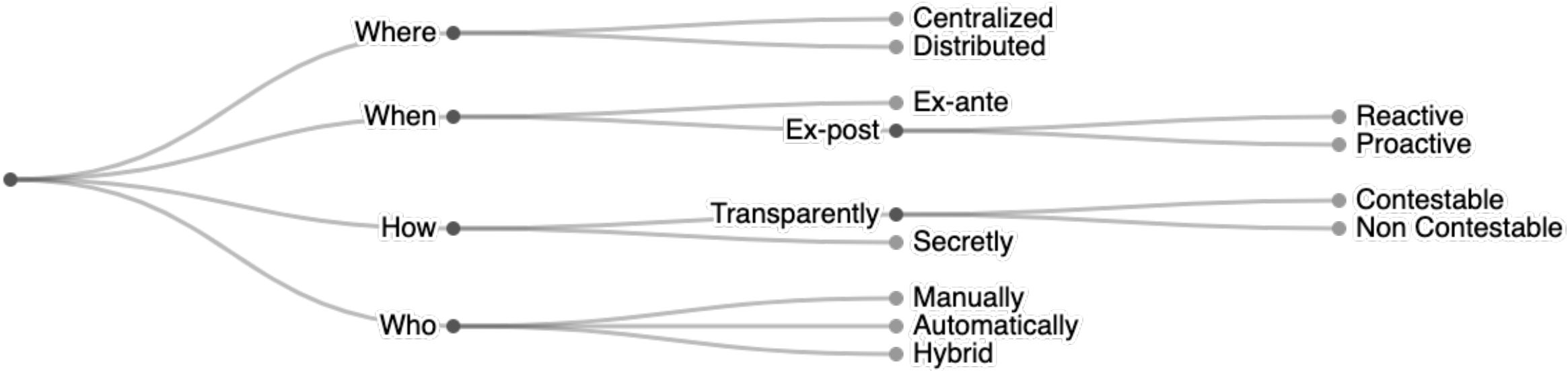
# What is moderation?

- 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


Where
When
How
Who
Centralized
Distributed
Ex-ante
Ex-post
Reactive
Proactive
Transparently
Secretly
Contestable
Non Contestable
Manually
Automatically
Hybrid

# 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 signaled 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.

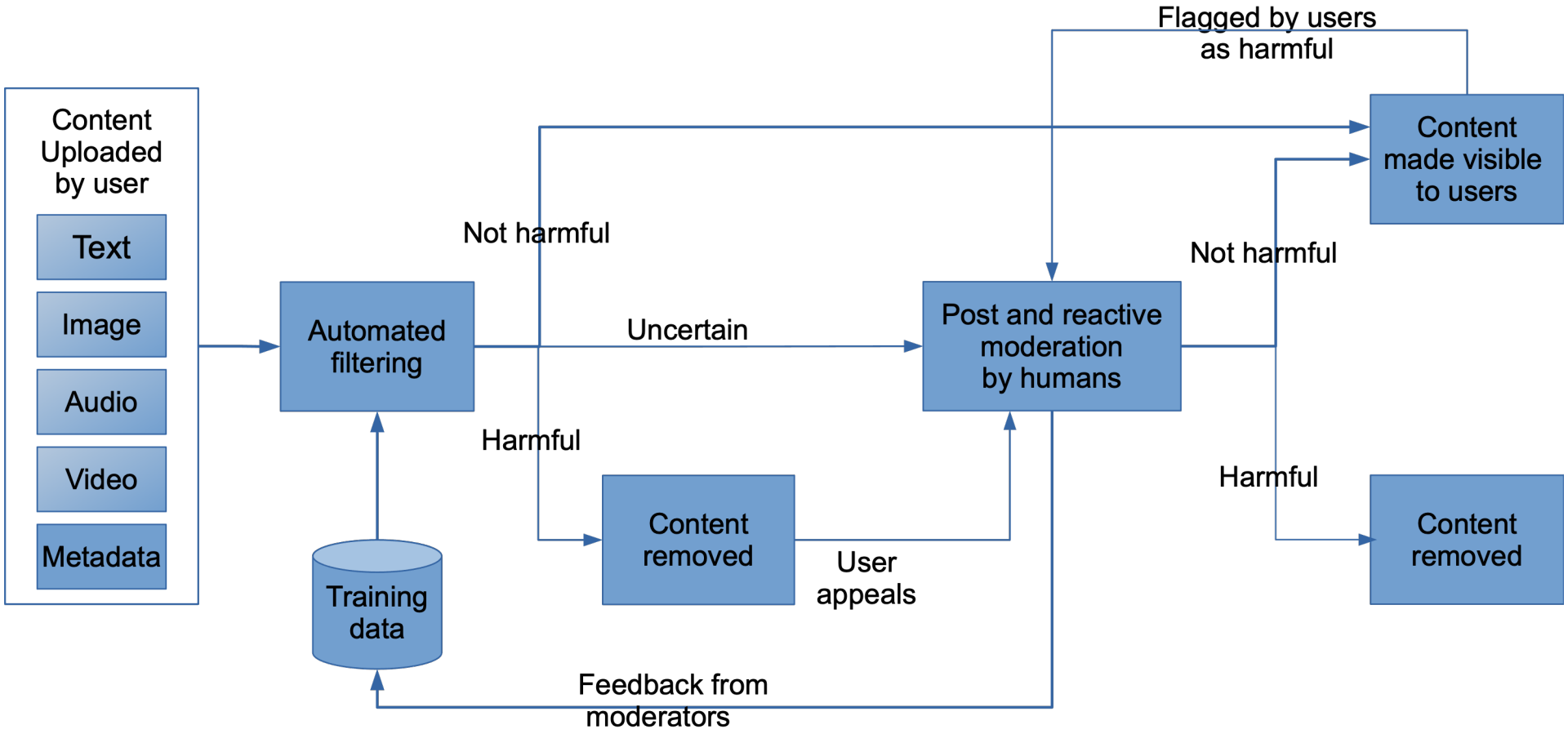
# Taxonomy - Who

- *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


Content Uploaded by user
Text
Image
Audio
Video
Metadata
Automated filtering
Not harmful
Uncertain
Harmful
Flagged by users as harmful
Content made visible to users
Post and reactive moderation by humans
Not harmful
Harmful
Content removed
Content removed
User appeals
Training data
Feedback from moderators

# Epic Fails

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


## Facebook banned Neptune statue photo for being 'explicitly sexual'

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

# Epic Fails

WIRED BACKCHANNEL BUSINESS CULTURE GEAR MORE SIGN IN

ISSIE LAPOWSKY BUSINESS 03.15.2019 01:50 PM

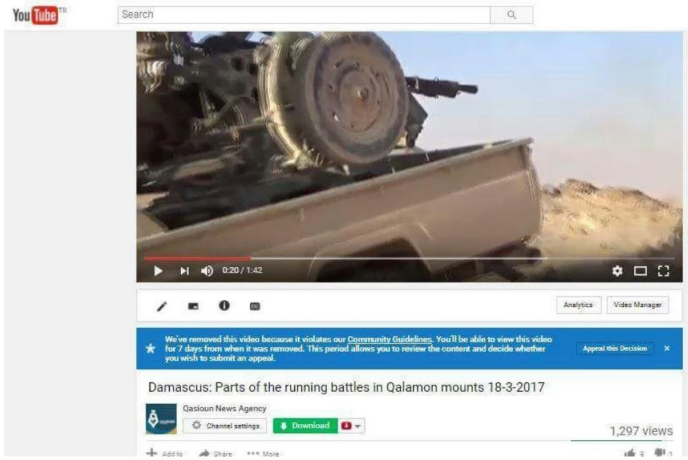
## Why Tech Didn't Stop the New Zealand Attack From Going Viral

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...

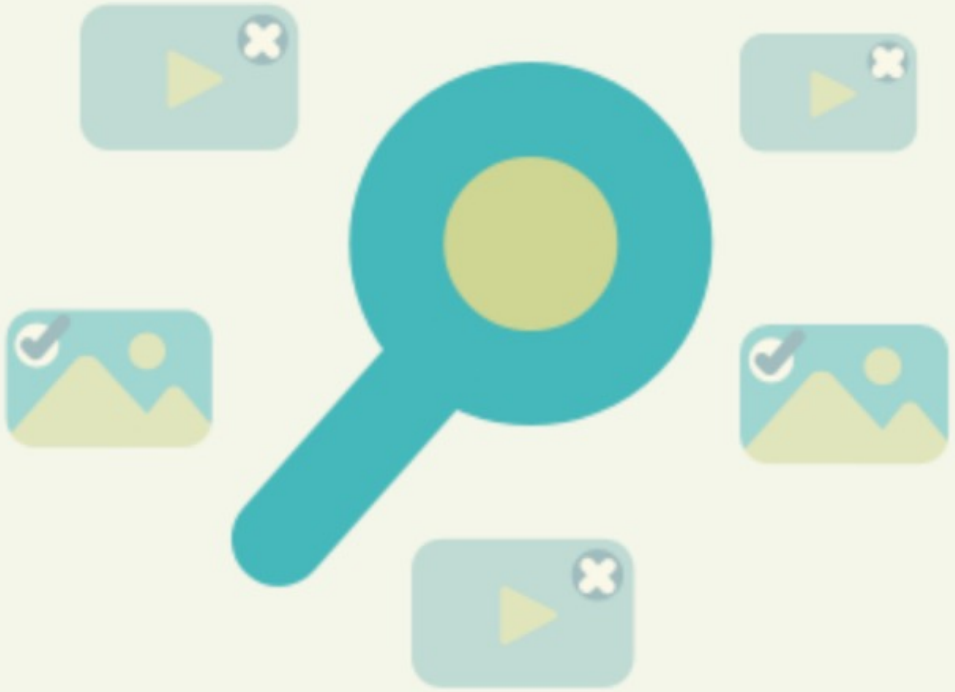
The New York Times

## YouTube Removes Videos Showing Atrocities in Syria




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

# Santa Clara Principles



**Companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines.**

# Santa Clara Principles

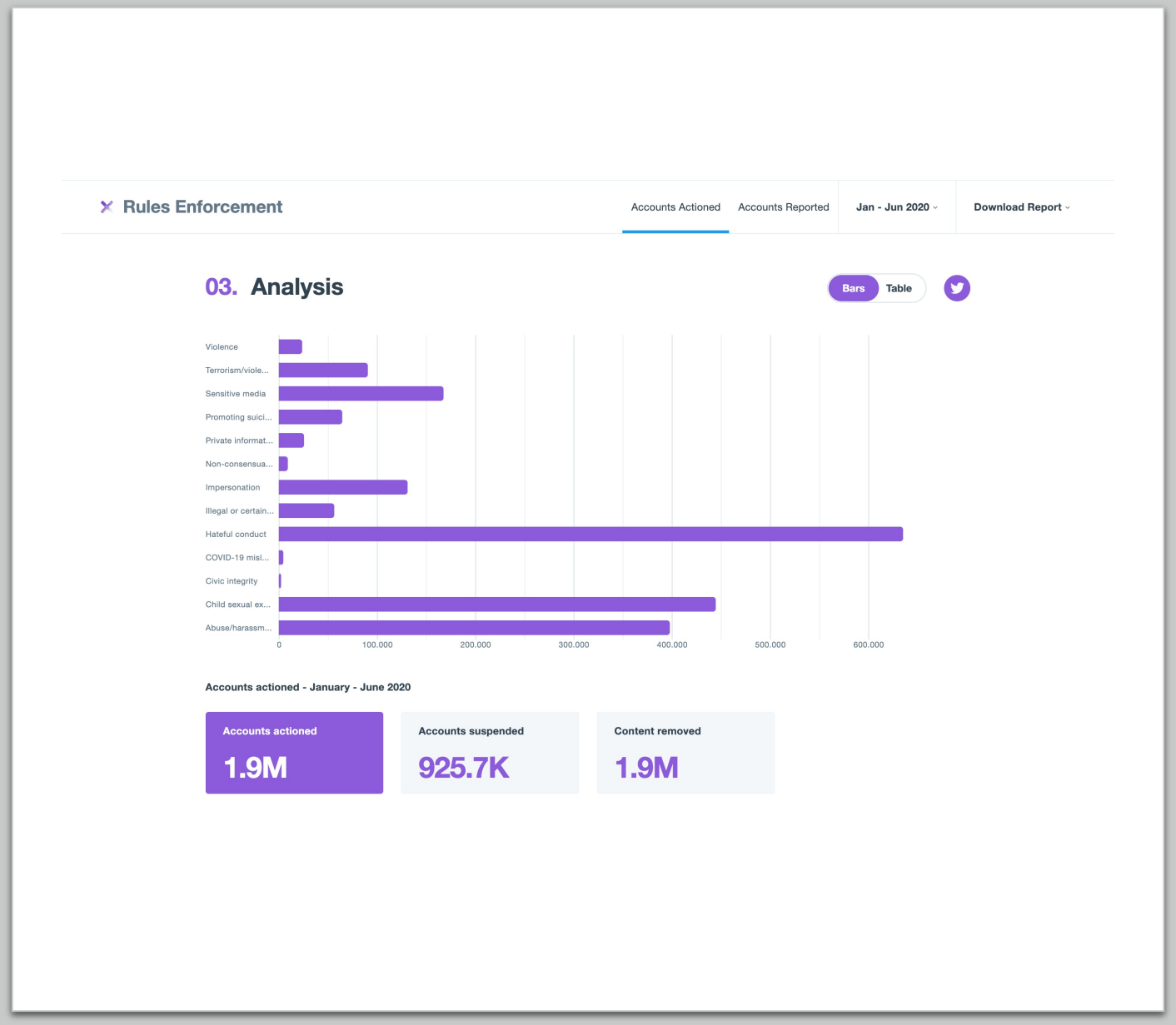
**Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension.**

# Santa Clara Principles

**Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.**

# Transparency

- Example from Twitter transparency report



# Issues on

- Filter bubbles
- Echo chambers
- Censorship
- Fake news

MAI4CAREU Master programmes in Artificial Intelligence 4 Careers in Europe

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423