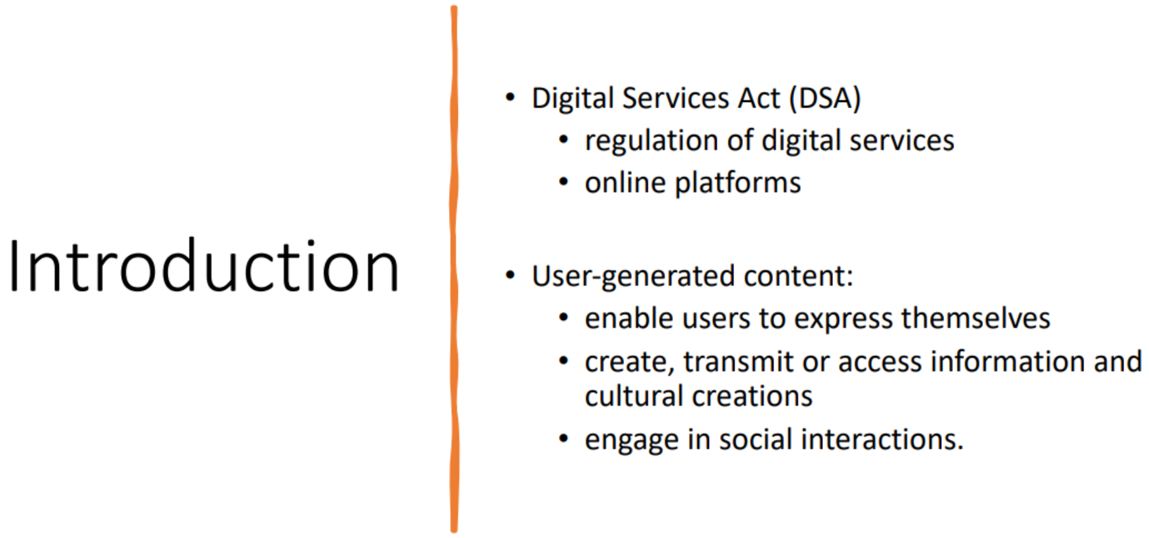
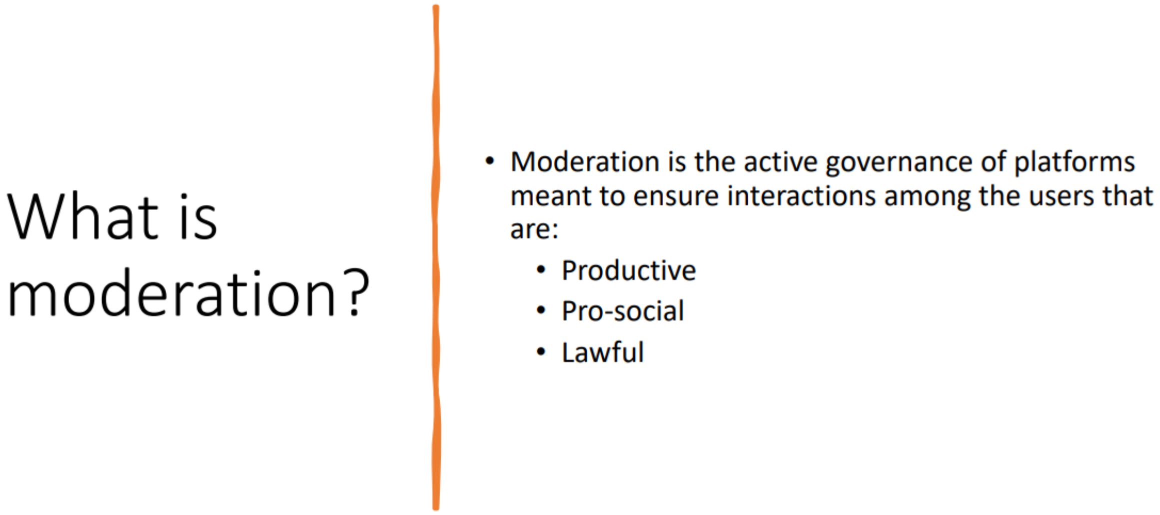
Lesson_13_Ethics_of_filtering



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.

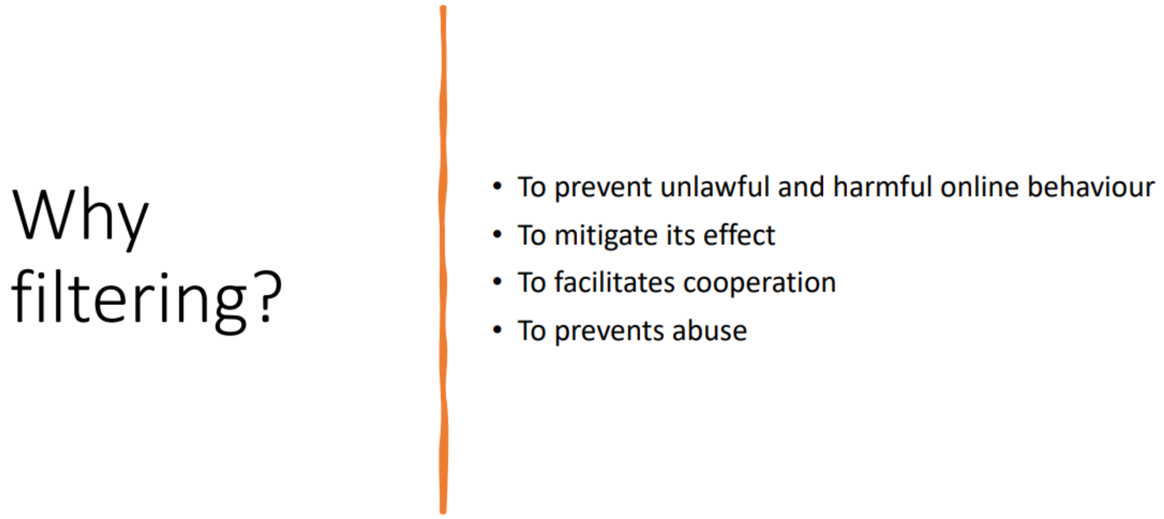
The DSA will replace the old directive about e-commerce and will try to regulate all those platforms that provide digital services and environments that are useful for the users to **exchange information** through user-generated content. We should bear in mind that UGC is a way for users to **self-express**.



What is moderation?

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

Moderation is *mandatory* in this kind of environments. It is a set of techniques used for an active governance of the platforms and ensure the users' interactions are *productive*, *pro-social* and *lawful* without any harm.

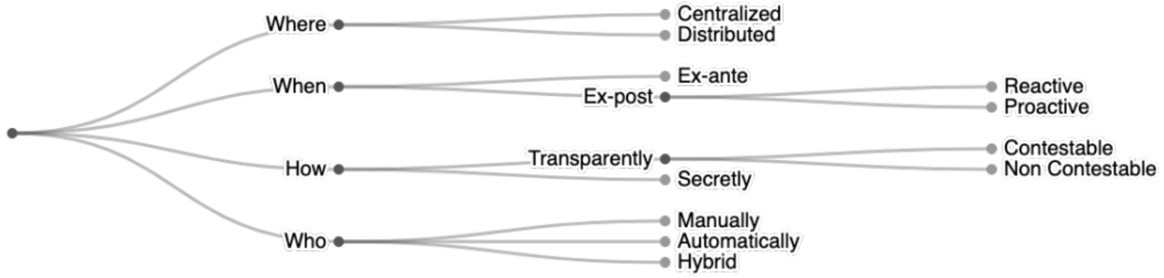


Why filtering?

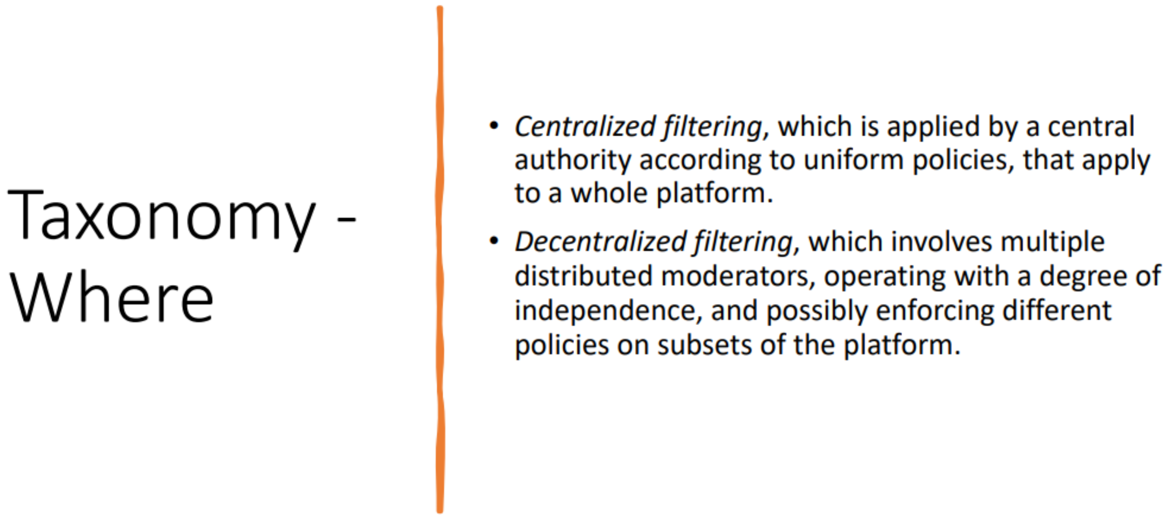
- To prevent unlawful and harmful online behaviour
- To mitigate its effect
- To facilitates cooperation
- To prevents abuse

Filtering is needed sometimes to prevent unlawful and harmful behaviours or mitigate possible damages that may derive from them. It is considered filtering any kind of technique that is used to ban or remove content from an online platform.

Taxonomy

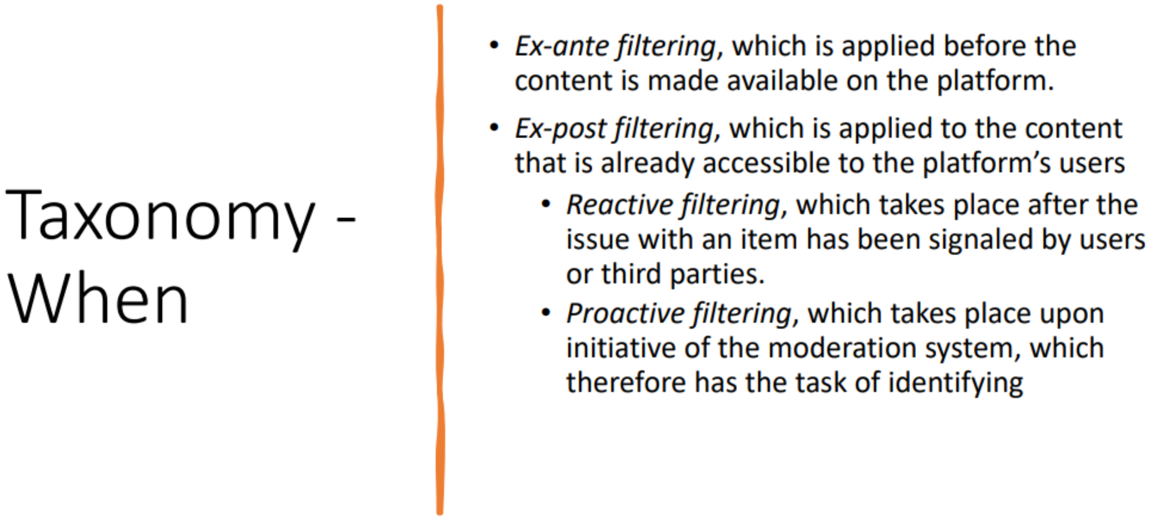


The above taxonomy describes the characteristics that distinguish different techniques that can be used to filter out contents.



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.



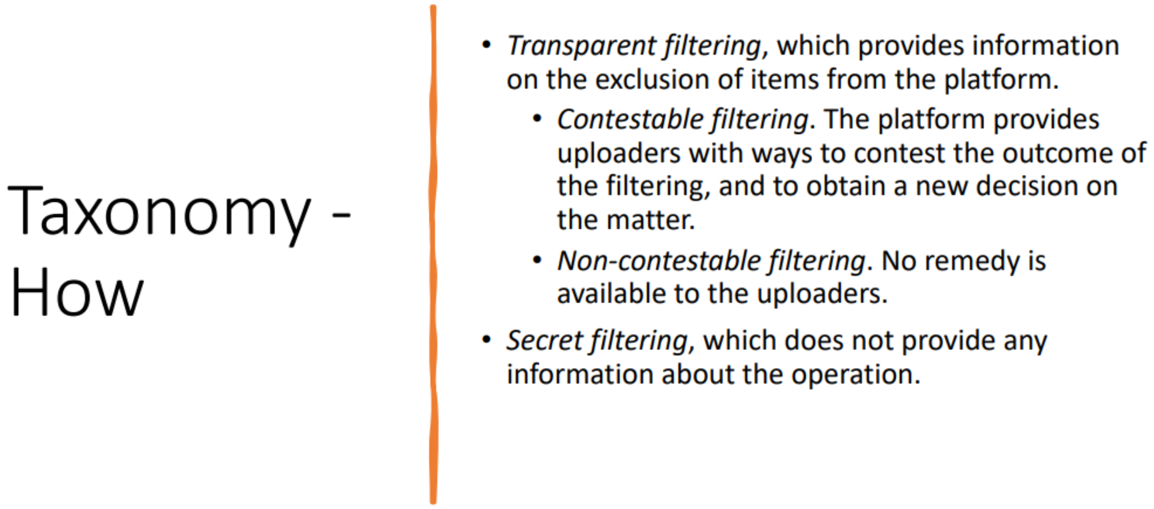
Taxonomy - When

- *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
  - *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

If the filtering action takes place before the content is published, we talk of ***ex-ante filtering***, or if the filtering is performed after the posting, we talk of ***ex-post filtering.***

The latter is a reaction that the platform can enact to spot any kind of illegitimate content and we can make a further distinction:

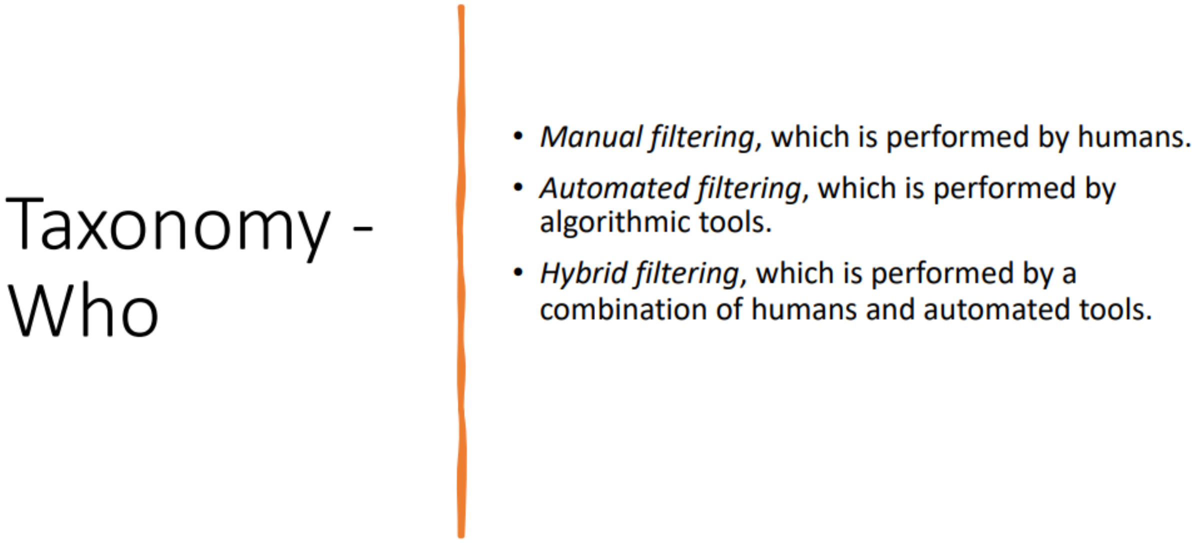
- **Reactive Filtering:** performed after the content is published and it **has been signaled** by the users, so the platform reacts and examines the content deciding whether it should be removed or moderated in some other way.
- **Proactive filtering:** the platform's policy could be that after the content is published is also checked to identify possibly illegitimate content.



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.

- ***Secret filtering:*** the user is not aware of the filtering action.
- ***Transparent filtering:*** the users are aware that exists some kind of filtering on the platform.
    - ***Contestable filtering:*** users are allowed to send messages to the platform and contest the decisions it made on the filtering.
    - ***Non-Contestable Filtering****:* the users cannot complain on the filtering action.



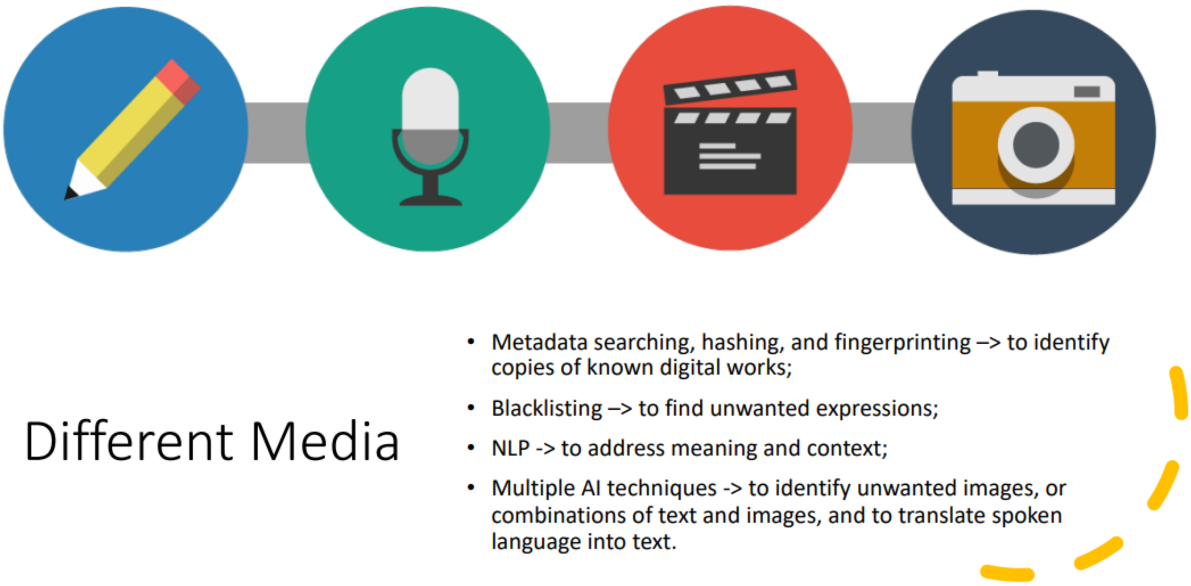
Taxonomy - Who

- *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.

Who applies the filters?

- ***Automatic filtering:*** some kind of AI technique examines the data/information and decides whether it should be filtered or not.

- ***Manual filtering:*** when the data are examined by human moderators.
- ***Hybrid filtering:*** combination of the two above. For example, we may have a first level of automated filtering that spots potential unlawful messages and and a second level of manual filtering that polishes and avoids possible errors performed by the first level .



Different Media

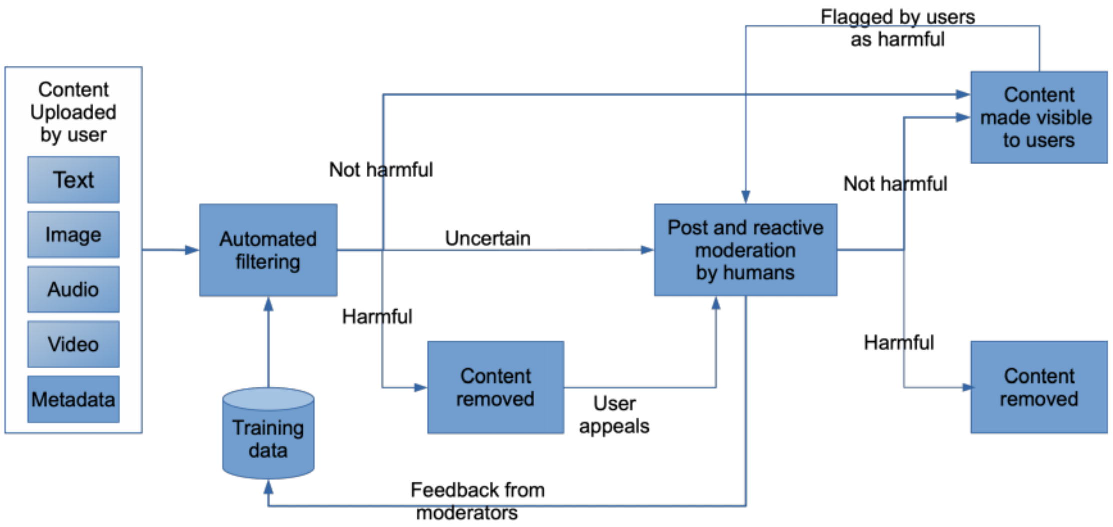
- Metadata searching, hashing, and fingerprinting –> to identify copies of known digital works;
- Blacklisting –> to find unwanted expressions;
- NLP -> to address meaning and context;
- Multiple AI techniques -> to identify unwanted images, or combinations of text and images, and to translate spoken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

We have different techniques, as we have seen above, but we have also different kinds of media. Diverse content can be shared on digital platforms by the users, and the type of content (text, audio, video etc.) changes the type of filters we are going to apply.

For example, hashing and fingerprinting use an algorithm to change the input's dimensionality adding a unique identifier which is compared with a database of non-abusive or copyrighted fingerprints, in this way we can identify unlawful or copyrighted content. This is the easiest way to check for unlawful content but is also very easy to deceive: small changes in the input can change the unique id, making the matching with the database impossible.

AI techniques are being used to overcome this kind of shortcomings, trying to understand the content of the uploads (NLP for text, DeepNN for images...), but they can be fooled too.

How it works



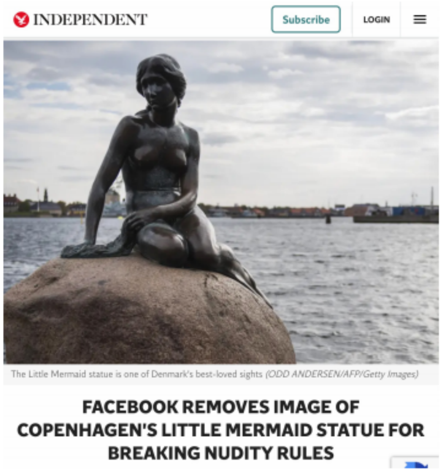
This is a high level schema of how this techniques work.

1. The user tryes to upload some kind of content.
2. There's a first level of **automated filtering** that tries to understand if the content can be problematic. It classifies the message in 3 different categories:
   - *Not harmful:* the message can be published, and user can flag in the case that the AF was mistaken (usually AF are based on ML approach, therefore prone to errors).
   - *Uncertain:* the input nature is not clear, a second level of human moderators kicks in and try to classify the input.
   - *Harmful:* the content is removed before being posted, and again the human team checks if the content is righfully classified.

All the feedback from the moderators is then inserted in the training data of the AF, to improve its performances.

Epic Fails

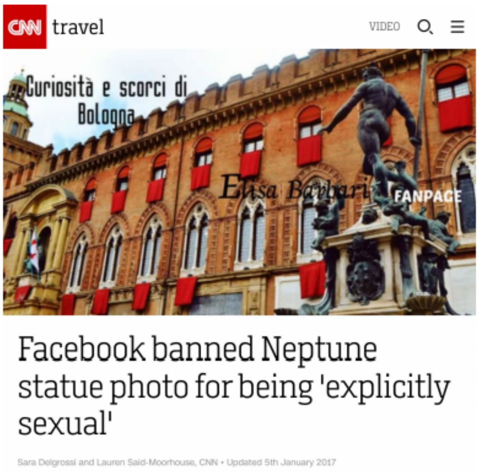
Epic Fails



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



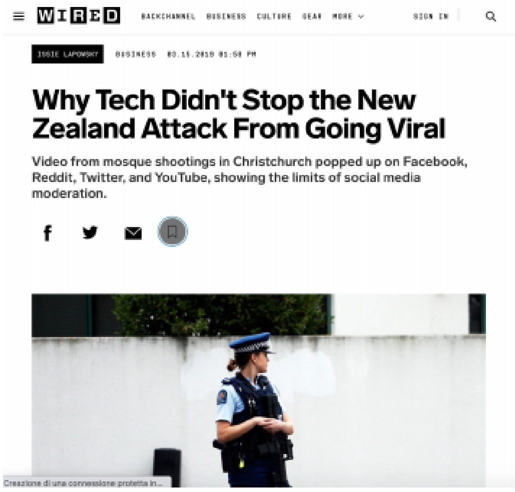
CNN travel VIDEO

Facebook banned Neptune statue photo for being 'explicitly sexual'

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

Artificial intelligence techiniques and algorithms ***lack of common sense***. For example, they will take the guideline of banning ANY images of nudes strictly, even if the image shouldn't be banned like in the case of statues.

Epic Fails



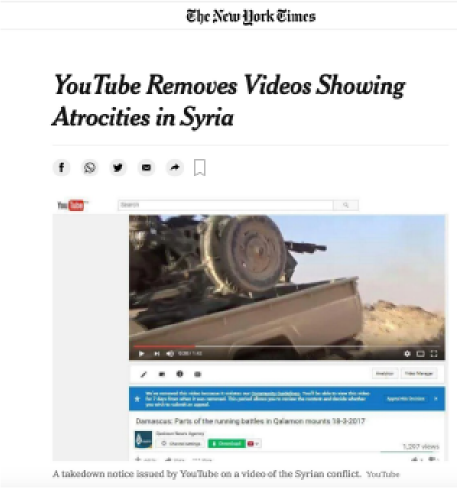
WIRED BACKCHANNEL BUSINESS CULTURE GEAR MORE SIGN IN

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

Damascus: Parts of the running battles in Qalamon mounts 18-2-2017

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

A video of a terrorist attack in New Zeland was broadcasted live and became viral on many different social networks. None of these platforms banned the content because their filters were not able to automatically recognize it as harmful and block it. On the other hand, youtube blocked in 2017 many videos of the civil war in Syria.

The content of these videos is very similar, so why is one banned (the wrong one) and the other isn't? [ideally we would want to ban the terrorist attack video and publish the on about the Syria] This is another example of machines having a hard time to distinguish the content, especially in videos where there are a lot of different media (audio,images..) and a lot of information. Also, the broadcast live leaves little time to analyze the content and decide whether it should be banned.

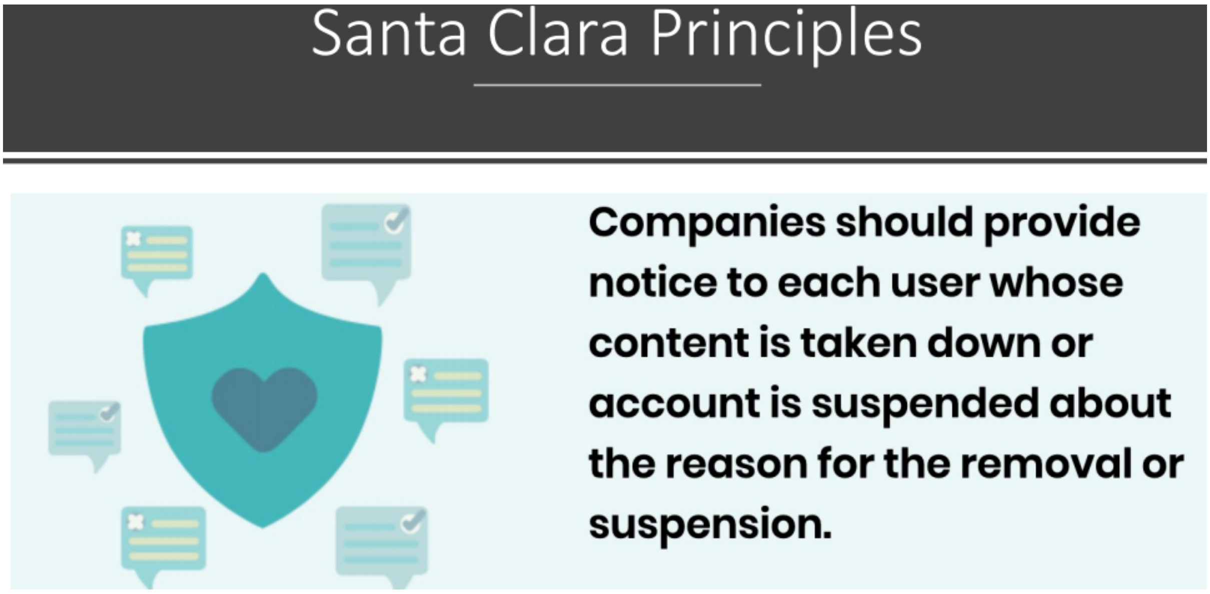
Towards Transparency...

As we said, these techniques might make some mistakes and the bigger the platform, the bigger the number of users that might be involved in those errors. So, there are many initiatives, Santa Clara among them, that try to push companies to be more trasparent and publish more information on how the filters work, so that the users can better understand what's going on. The Santa Clara principles are 3 and, if followed, should help to be more transparent.

Santa Clara Principles

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

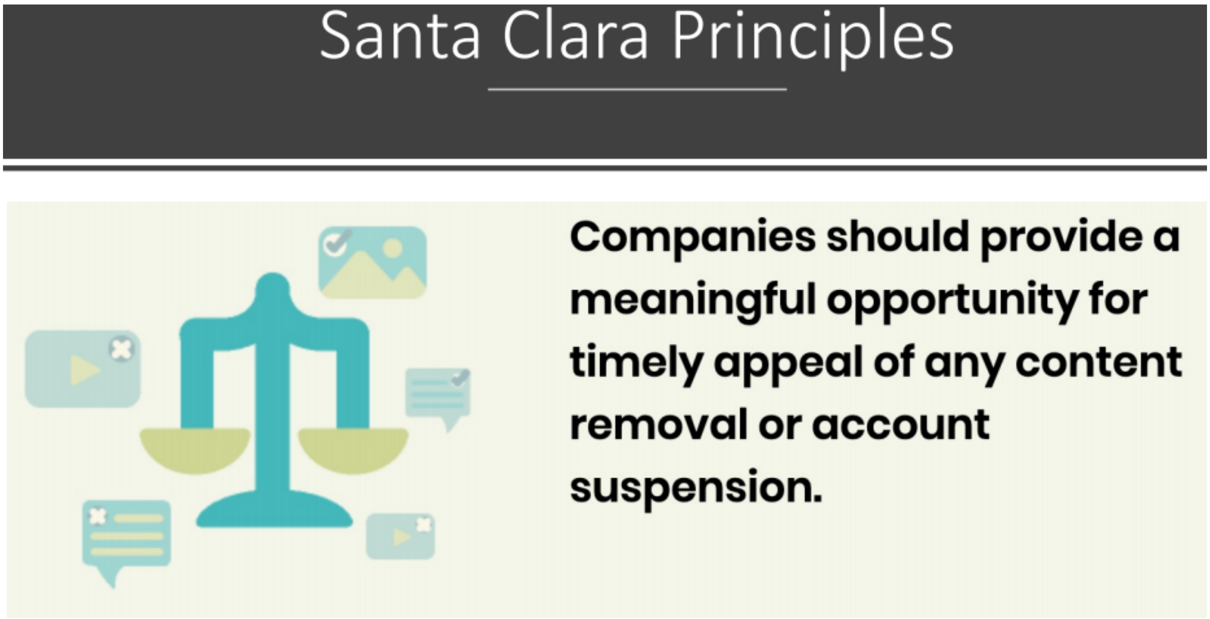
Company should be transparent on the quantity of content and users are stopping through the filtering.



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.

The company should make the users aware of the reasons why their contents are banned or stopped.



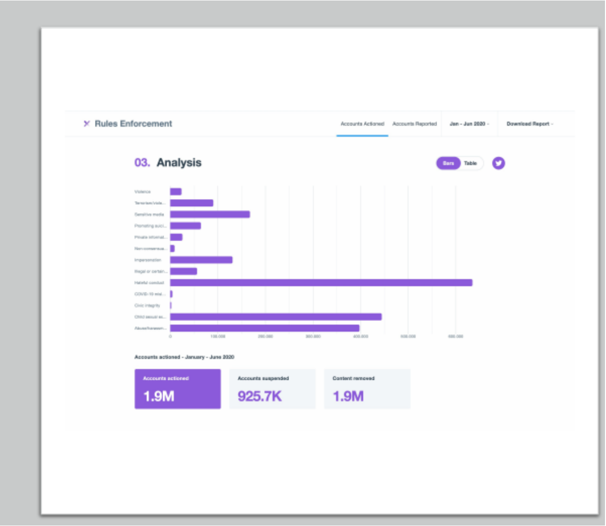
Santa Clara Principles

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

When users are notified about a possible suspension of their content they should have a way to ask to reconsider the decision and possibly public or at least have an explanation on why the content is stopped.

Transparency

- Example from Twitter transparency report



Rules Enforcement

03. Analysis

Accounts actioned - January - June 2020

Accounts actioned	Accounts suspended	Content removed
1.9M	925.7K	1.9M

This initiative started to design a transparency report which describes the statistics of the different categories affected by the filtering procedure.

Issues on

- Filter bubbles
- Echo chambers
- Censorship
- Fake news

There are several issues related to this technology that may influence the users' information elaboration.

- ***Echo chambers & filter bubbles***: the filtering creates an environment in which opinions become fragmented and for users is more difficult to process information or to change their minds, since most of them think in the same way within that specific environment.
- ***Censorship***: finding a the right level of censorship is a key point to have the right moderation without violating the users' rights.
- ***Fake news:*** with the technological progress, specifically GANs, it has become harder to tell apart fake content from "natural" content and classify fake news correctly (e.g., DeepFake).