Information quality & moderation
Seminar by Google, 24. November 2020

Presenters:
[redacted]

Agenda:
- Google mission & key prinicples
- Google's Strategy - 4Rs (remove, raise, reduce, reward)
- COVID-19 case study
- Collaboration
- Digital Services Act

Mission & Key principles
- Main mission is **to organize the world's information and make it universally accessible and useful**.
- In that approach, the key focus of Google is on relevance and quality of the information
- Google offers a variety of services and products in a constantly changing environment – however, their mission and key principles are valid for each of those: Google wants to provide the best information at the right time and limit the impact and reach of low-quality information.
- Google has different levels of restrictions for each service. **Moderation and removal depend on the particular service**.
    - Search shows millions of pages and organizes the internet. Google believes that **users should find every legal result on their query**. Moderation and removal is therefore quite limited on Search.
    - The other end of the spectrum is Google Ads. Google's advertising products are the most strictest moderated.
    - Gmail also involves minimum levels of content moderation.
    - Youtube requires a broader restriction approach due to the high user exposure.

Google policy of 4Rs

1. **Remove**
    - General points
        - Google values openness and accessibility – but that doesn't mean anything goes
        - Generally, if a content is not forbidden by law or runs contrary to community Guidelines, then it should be findable. Google leans to keep such content accessible and respect user choices.
    - Approach
        - Google takes its responsibility very seriously and **sets rules for all its services. They take action based on these rules and local legal obligations.**
        - Harmful, but not illegal content, is not removed but reduced (see below) to limit its spread.
        - Notification are assessed carefully and removals conducted on a case-by-case basis.
        - There is a possibility to challenge a removal – reviews after a complaint about a removal are always handled by human reviewers directly (unless it is content that has already been flagged and complained about previously).
    - Content removal on Youtube
        - The number one reason Youtube removes content is child safety, followed by spam and misleading content
        - Google just published the latest quarterly transparency report
    - Design of moderation policy

    - Google considers diversity when removal and moderation policies are developed.
    - Safety and quality information is at the core of the design of new policies. Trusted flaggers and regulators are frequently consulted for their expertise.
    - Then Google gathers as many examples as possible on how the specific harm might manifest itself in the future.
    - With that information a standards is drafted.
    - **Again, Google internal experts are consulted and a high level of consistency is ensured, before any new policies are enforced.**
    - Users are then informed about any potential changes
- Use of automated technology
    - Automated technology is used to flag content en masse before users see it. Trained reviewers help to assess context-heavy or new content that has never been seen and is not interpretable for technology.
    - Google carefully balances the use of automated technology as they are aware on the risks of over-reliance on machines.
    - Two types of automated technology are used – machine learning and hashing
    - For **machine learning**, algorithms get indicated what content is violative. Google does not instruct algorithms to search for something in particular. Instead, they highlight what is being removed and the algorithm learns what elements are undesirable based on the removals and act based on this learning.
    - For **Hashing**, this type of automated technology aims to catch re-uploads of known violative content. To do so they use a “digital fingerprint” that every Youtube-Video has
- Success of automated technology
    - One successful example is violent extremist content**.**
    - When Google started with automatic recognition in Q 2017 only 8 per cent was removed before the content hat 10 views
    - **In Q4 2019 the rate was at 90 per cent. T**
    - The technology allows Google to clean the platform at scale.
    - Upon question: in some areas like violent extremist content, automated technology is good and reliable. In other areas, this is not necessarily the case and a lot of the content needs to be reinstated after review and complaints
- Need for human reviewers
    - Human reviewers are critical to evaluate context and to establish the intention of the upload. Why was the content in question uploaded, what intention? Has it a documentary value?
    - For ex. a graphic footage from a warzone. The same content can be used by a terrorist or a civil rights organization. This is why context matters so much.
    - The same applies to hate speech. Here you have very easy examples. For ex. a Swastika. A machine could easily recognize this as hate speech. But then: You have American History X, an excellent movie or old Buddha statue sporting a swastika. A machine can hardly understand the context.
- Challenges for human reviewers
    - For content moderators a lot of content can be disturbing. **Google uses machines to blur or black and white some disturbing pictures.**
    - Google has invested significantly in that area.
- Transparency on Government removals
    - Google discloses requests by governments.

2. Raise

- **Google elevates high quality and trustworthy information and** revents the spread of harmful content.
- They use algorithms to organize content on their usefulness. For this, the interest of users is first established by whatever is put in a searchbar or watched (e.g. football match)

- Ranking systems, we use ranking algorithms quality. For ex. Users are looking for specific content or advise.
- On authoritative sources: In some cases it is quite straightforward (e.g. WHO with COVID), in others it is much more complicated and Google has to balance it carefully. Google also relies on fact-checking agencies to establish authoritativeness.

3. Reduce

- Aim here is to reduce the spread of potentially harmful content and not to proactively expose users to harmful content. Example here is the autocomplete feature and a proactive removal of popular terms that are harmful
- For recommendation, Google has historically looked three signals:
    - What have you looked at before?
    - What do people enjoy who have watched similar content that you did?
    - Popularity of videos
- Three changes have happened
    - Recommendation of click-bait is reduced
    - User satisfaction – bar with regard to whether you enjoyed video or not at the end of a video
    - Promotion of authoritative content

4. Reward

- Google doesn't want anyone to make money on bad content. They take action against bad advertising as well. In 2019 more than 2 mio bad ads were removed
- Making money from ads is only possible for good content creators and is a privilege that must be earned.

<u>COVID 19 Case study</u>

- Since Jan 2020:
    - Removed over 82.5 Million coronavirus-related ads
    - Since February, over 600'000 videos have been removed from Youtube due to Covid-lies (over 96% have been removed automated)
    - Highlighted fact-checks and links by health authorities. Overall, fact-checks appear 6 Mio times on average. That us why fact-checking has also been rolled out to Youtube.
- Reducing the reach of borderline content and misleading political speech linked to Covid at the same time
- Helping others raise awareness about Covid-19 disinformation via Google News – to help debunk common and misleading Covid-19 claims

<u>Collaboration</u>

- Google is member of the technology coalition fighting child abuse
- Another example is GIFCIT. Establish a global internet forum to counter terrorism.

<u>Digital Services Act</u>

- Policy considerations by Google
    - Preservation of country of origin principle
    - Liability only for clearly-defined illegal content
        - Harmful content that is lawful should be addressed by self- and co-regulation.
    - No general monitoring obligations
    - Harmonised, graduated liability scheme depending on service type
        - Greater responsibility comes with greater control of the content.
        - But different services have a different outlook and context.

        - Google would propose different services categorization than in eCD.
    - Same rules for all players, no matter their size
    - Meaningful transparency and oversight
    - Notice and action
        - Need for harmonised and strengthened notice formalities
        - Risks posed to EU fundamental rights by mandatory short TaTs and staydown obligations
    - Removing disincentives for voluntary actions
        - Idea of “active” service provider runs risk of losing liability safe harbour
        - Need of clarification to make sure that intermediary should be able to manually and voluntarily review content for specific unlawfulness (e.g. illegal terrorist content) without risking losing the safe harbour for other unlawfulness (e.g. hate speech)

Q & A

- Courts play a crucial role, the highest authorities, when it comes to difficult decisions. **Awful content that is not illegal is simply been downrated but not removed. We have to respect Freedom of expression. Transparency reports on political ads.**
- The improvements of the technology. We are not in a situation where tech is 100 per cent clear and effective. But we see the potential. **The machines are good for some types of content but not for others. Context heavy content is difficult to rate.** I cannot tell you where the machines are going, it is really hard ,we train a lot of systems. But we launch it only if they are reliable enough. It has to work on different types of content, video, text, audio, you have to train differently. Also for the languages. Hate speech in France, in French-Canada. The more data you have the more efficient you are.
- A lot of video is very dynamic. Machines need to learn, it takes a lot of data and therefore it takes a lot of time.
- **We are looking at different metrics regarding quality content. It is also location specific.** We want to provide timely information. The COVID 19 has shown us, how important it is to have good medical information. We do have local health authorities. We can say this is the information where we can check the content. If there is information which goes against WHO guidelines it is much easier to remove. **This gets more difficult when it goes to political speech and propaganda. Indeed there are ways where we can measue it, factcheckers help us to give users additional context.**
- An example of borderline content? There was a documentary called older with a lot of difficult information propagating conspiracies. We downrated this but not remove it.
- We want remain open how to bring in the experts. This seminar is also a way how we want to strengthen our collaboration with audiovisual regulators.
- **Each time we launch new policies, we consult with civil organisations and academic experts.** Content that is extremely difficult to judge what we do, we reach outside. We have ways to consult.