COUNTER EXTREMISM PROJECT

*Policy Brief*

# Terrorist Content Online -
## How to build comprehensible transparency for automated decision-making systems (ADM)

Alexander Ritzmann
Prof. Hany Farid

The dissemination of terrorist content is one of the most widespread and most dangerous forms of misuse of online services[1]. The current reporting mechanisms on preventing the dissemination of terrorist content, however, do not provide enough data or information to properly understand how social media platforms are being used by terror-groups. More transparency is therefore required to allow policy makers and civil society to understand how social media platforms are being weaponized against society and democracies.

An estimated 720,000 hours of video content are uploaded to YouTube every day, and some one billion posts, including 300 million images, are shared on Facebook each day. To process this amount of data, social media companies already apply upload and re-upload filters to keep illegal or unwanted content off their platforms.

Reservations against proactive measures and automated decision-making systems are understandable. Yet, the question is no longer *if* (upload-)filters should be applied to prevent the dissemination of terrorist content online, but *how* to apply them. Smart regulation that focuses on transparency, explainability and effectiveness will protect civil liberties more than no regulation.

The “Ethics Guidelines for Trustworthy Artificial Intelligence” of the “EU High-Level Expert Group on AI” highlight the importance of transparency and explainability of automated decision-making systems that have significant impact on people’s lives[2]. Such transparency would also lead to more accountability and would allow regulators to apply sanctions when appropriate.

An essential part of transparency is the ability to explain both the technical processes of the applied ADM-systems and the related human decisions. A more transparent reporting mechanism must therefore include explanations of the individual automated moderation tools, the technical compliance system as a whole as well as a better understanding of the application of moderation policies in practice.

[1] https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/649326/EPRS_BRI(2020)649326_EN.pdf
[2] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

COUNTER EXTREMISM PROJECT

An appropriately transparent system requires two main features to be effective. First, a suitable entity, with the appropriate technical and domain expertise, should be designated as an external observer with full access to moderation policies and procedures.
Second, transparency reports must provide detailed and comprehensible information as enumerated below. The designated entity may limit the level of detail of the published reports to protect trade secrets and to prevent misuse by criminals.

| **15 main features of comprehensible transparency and questions to be addressed in a transparency report** |
|---|
| 1) What are the underlying "theories of change" and theoretical concepts for the moderation tools and systems? |
| 2) What classification criteria are used to search for content? |
| 3) What is considered "terrorist", "extremist", or "illegal" content? |
| 4) Which content categories (e.g., text, images, videos) are being searched and classified? |
| 5) Which AI or machine-learning systems are being applied for content moderation? What is the accuracy of these systems? |
| 6) How is machine-based training data validated to avoid bias? |
| 7) What quality assurance or evaluation procedures are used? |
| 8) To what extent and in what function are human moderators involved? Which processes are in place to account for potential moderation bias? |
| 9) How many notices are received through users or trusted third parties? |
| 10) How many posts were detected by automated systems? |
| 11) How long, from the time a report is filed, did it take to block/remove content or decide not to block/remove content? How long does it take to inform all parties involved? |
| 12) Of all notices received, what percent of content was blocked/removed? |
| 13) Of all notices received, what percent are duplicates from previous reports (re-uploads)? |
| 14) Of all notices received, how many views did each posting/file receive before takedown? |
| 15) How is the well-being of human moderators monitored and addressed? |

The CEP Policy Paper "NetzDG 2.0 - Recommendations for the amendment of the German Network Enforcement Act (NetzDG)", April 2020, can be downloaded here: https://bit.ly/2S9L0Sd

For further information please contact us: berlin@counterextremism.com