COUNTER EXTREMISM PROJECT

CEP Policy Brief

Terroristische Inhalte Online

Grundlagen für nachvollziehbare Transparenz beim Einsatz automatisierter Entscheidungssysteme (ADM-Systeme) auf Social Media-Plattformen

Alexander Ritzmann
Prof. Dr. Hany Farid

Die Verbreitung terroristischer Inhalte ist eine der gefährlichsten Formen des Missbrauchs von Online-Diensten[1]. Die derzeitigen Transparenzberichte der Social Media-Unternehmen liefern jedoch nicht genügend Informationen um vollständig nachvollziehen zu können, wie genau Plattformen von Terrorgruppen aktuell missbraucht werden und wie dies zukünftig verhindert werden kann. Damit politische Entscheidungsträger und die Zivilgesellschaft besser nachvollziehen können, wie Social Media-Plattformen gegen unsere offene Gesellschaft und die freiheitlich-demokratische Grundordnung eingesetzt werden, ist eine erklärbare Transparenz der eingesetzten Moderationssysteme erforderlich.

Vorbehalte gegen proaktive Maßnahmen wie Upload-Filter und andere automatisierte Entscheidungssysteme (ADM) sind verständlich. Tatsache ist jedoch, dass schätzungsweise 720.000 Stunden Videoinhalt täglich auf YouTube hochgeladen werden, und rund eine Milliarde Beiträge, darunter 300 Millionen Bilder, auf Facebook innerhalb von 24 Stunden geteilt werden. Um illegale oder unerwünschte Inhalte von ihren Plattformen fernzuhalten, wenden Social Media-Unternehmen bereits seit Jahren Upload- und Re-Upload-Filter an. Die Frage ist deshalb nicht mehr, *ob* Filter eingesetzt werden sollten, um die Verbreitung terroristischer Inhalte im Internet zu verhindern, sondern *wie* sie angewendet werden sollen. Eine tatsächliche Transparenz und Nachvollziehbarkeit schaffende Regelung im NetzDG 2.0[2] ist deshalb dringend erforderlich, um die Bürgerrechte der Nutzer*innen zu schützen.

Die „Ethics Guidelines for Trustworthy Artificial Intelligence" der „EU High-Level Expert Group on AI" betonen die Bedeutung von Transparenz und Erklärbarkeit von ADM-Systemen, die erhebliche Auswirkungen auf das Leben von Menschen haben[3]. Eine solche Transparenz muss auch Verantwortlichkeiten definieren und es den Aufsichtsbehörden ermöglichen, gegebenenfalls rechtliche und finanzielle Sanktionen anzuwenden. Ein wesentlicher Bestandteil von erklärbarer Transparenz ist die Fähigkeit, sowohl die technischen Prozesse der angewandten ADM-Systeme als auch die damit verbundenen menschlichen Entscheidungen, nachvollziehen zu können. Eine tatsächliche Transparenz schaffende Berichterstattung durch die Social Media-Plattformen muss daher nachvollziehbare Erläuterungen zum Content-Moderationssystem als Ganzem, den einzelnen ADM-Instrumenten sowie der praktischen Umsetzung von Moderationsrichtlinien beinhalten.

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

Hauptmerkmale nachvollziehbarer Transparenz

Ein nachvollziehbares und wirksames Transparenz-Regime besteht aus zwei Komponenten. Erstens sollte eine geeignete Institution mit der entsprechenden fachlichen und technischen Kompetenz als externe Aufsicht mit uneingeschränktem Zugang zu Moderationsrichtlinien und -verfahren benannt werden. Zweitens müssen Transparenzberichte detaillierte und vor allem nachvollziehbare Informationen enthalten, die im Folgenden aufgeführt werden. Die mit der Aufsicht beauftragte Institution kann den Detaillierungsgrad der veröffentlichten Berichte einschränken, um Geschäftsgeheimnisse zu schützen oder den Missbrauch zu verhindern.

15 Fragen, **die in einem Transparenzbericht beantwortet werden sollten**
1) Auf welchen relevanten theoretischen Konzepten (z.B. „Theory of Change") basieren die angewandten Content-Moderationssysteme und -instrumente?
2) Wie werden „terroristische" oder „illegale" Inhalte definiert?
3) Welche Klassifikationskriterien werden für die Suche nach illegalen Inhalten verwendet?
4) Welche Inhaltskategorien (z.B. Text, Bilder, Videos) werden durchsucht und klassifiziert?
5) Welche ADM-Systeme („reinforcement learning"/„deep learning") werden zur Moderation von illegalen Inhalten eingesetzt? Wie hoch ist die Genauigkeit dieser Systeme?
6) Wie werden Trainingsdatensätze validiert, um (unbewusste) Vorurteile (Biases) zu identifizieren und zu vermeiden?
7) Wie viele illegale Inhalte wurden durch ADM-Systeme erkannt?
8) In welchem Umfang und in welcher Rolle sind menschliche Moderator*innen beteiligt? Welche Verfahren werden angewandt, um (unbewusste) Vorurteile (Biases) bei der Moderation zu identifizieren und auszugleichen?
9) Wie wird das psychische Wohlergehen der menschlichen Moderator*innen sichergestellt?
10) Wie viele Meldungen (Notices) gehen über Nutzer*innen oder „vertrauenswürdige Dritte" ein?
11) Wie lange dauerte es vom Zeitpunkt der Meldung (Notice) an, Inhalte zu sperren/entfernen oder zu entscheiden, dies nicht zu tun? Wie lange dauert es, bis alle beteiligten Parteien informiert worden sind?
12) Wie viel Prozent aller gemeldeten/gefundenen Inhalte wurden gesperrt/entfernt (unterteilt nach ADM/menschlicher Moderation/Nutzer*innen)?
13) Wie viele Clicks/Views/Shares/Likes haben die Inhalte vor der Blockierung/Entfernung erhalten?
14) Wie viel Prozent aller gemeldeten Inhalte sind Duplikate von vormals gesperrten/ entfernten Inhalten (re-uploads)?
15) Welche Qualitätssicherungs- oder Evaluationsverfahren werden angewandt?

Mehr zum Thema: CEP POLICY PAPER - NetzDG 2.0: Empfehlungen zur Weiterentwicklung des Netzwerkdurchsetzungsgesetzes (NetzDG)