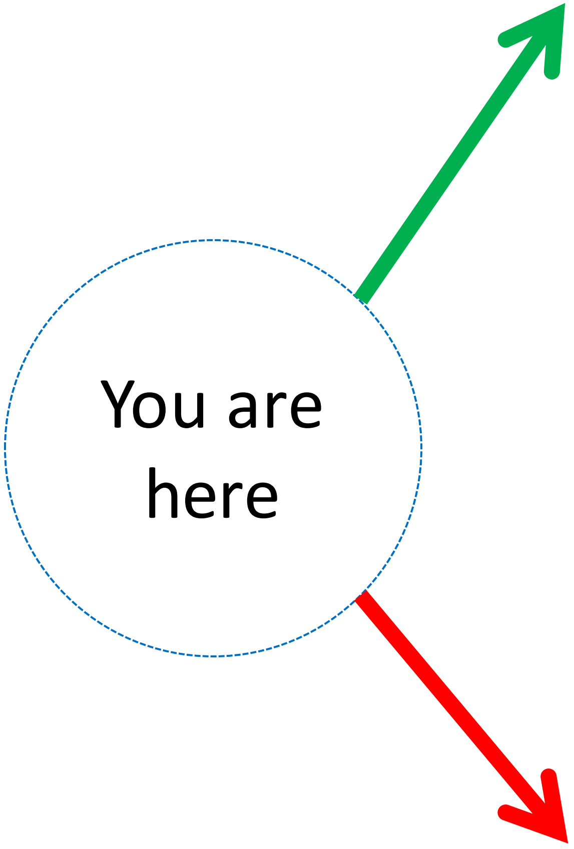
The role of AI in regulating abuses of social media

Prof. Anthony Clayton, CD
Chairman

BROADCASTING COMMISSION

Artificial Intelligence for Information Accessibility
AI4IA
UNESCO
28th September 2020



You are here

Transition to a digital economy opens up vast array of new opportunities.

Threatened by growth of organized crime and terrorism in cyber-space, use of social media for undermining democracy, encouraging extremism & violence, destroying trust in society.

Terrorism and social media

- IS used social media to control narrative, broadcast propaganda and atrocities, recruit disaffected youth across the world. By 2015 they were making up to 200,000 posts per day; recruited 40,000 foreign nationals from 110 countries to fight in Syria.
- ..and disseminate information. EU reported 54,000 websites with information on IEDs posted online by IS August 2016 to May 2017; 2/3rd of information shared in two hours of posting.

Christchurch Massacre 15th March 2019

- Brenton Tarrant killed 51, injured 49, live-streamed the massacre on Facebook.
- In the first 24 hours, YouTube had one upload per second on their platform. Facebook removed 1.5 million videos and blocked another 1.2 million at the point of upload. But the material was still there six months later.
- Attack then copied by others in e.g. Norway, Thailand, Ireland, USA

The UK’s National Coordinator for the Prevent counter-extremism program, warned that terrorists who self-radicalize using online material are now far greater threat to the UK, in terms of volume, than those who are directed and mobilized by a terrorist organization abroad. He said ‘an explosion’ of material inciting violence was accessible, and that young and vulnerable people and those with mental health issues were being targeted and exploited.

The volume of media content is beyond human ability to screen

Every minute:

- Over 500 hours of video uploaded to YouTube
- Nearly 250,000 photos uploaded to Facebook
- Over 500,000 comments posted on Facebook
- 30 million WhatsApp messages sent

It would take a lifetime to view the content uploaded to YouTube every day.

Can regulation be automated?

Algorithms can be used to search. But there are still problems:

- ☒ Bias. Algorithms search existing data; and can pick up biases in the data set used for training; e.g. crime predictive algorithm that learns from police reports will reflect any prejudices in those reports.
- ☒ Context (e.g. words modified by context or intonation).
- ☒ Language evolves (especially street language).
- ☒ Misinformation can be disguised e.g. spurious information about vaccines presented in pseudo-scientific manner that makes it appear credible.

Can regulation be automated?

- So algorithms can reduce volume, but cannot replace humans in final rounds of screening.

- Solution is likely to involve a combination of improved algorithms and tiered human screening.

- Emphasize prevention: promotion of digital literacy allows people to utilize digital resources while guarding against malicious content.

Legislative challenges

- Most existing international law on terrorism does not apply to cyber-attacks as harm is not physical but ‘merely’ disruptive.
- Can be hard to identify perpetrator (are Russian trolls independent or state agents?).
- Most regulators can’t demand that companies abroad remove content as they don’t have jurisdiction.
- Strong encryption makes criminal and terrorist activity invisible to law enforcement, but is also important to e.g. dissenters in totalitarian states.
- Deleting ‘suspicious’ content that does not have an unambiguous link to violence could be portrayed as censorship, will support extremist narrative of persecution.

Future direction?

- Probably towards greater supervision of cyberspace, relying largely on automated content monitoring systems and giving additional legal responsibilities to social media and technology companies.
- Challenge is to find ways to limit harms being caused while protecting democracy, freedom of expression and level of personal privacy.
- The solution will be hybrid, combining regulation, sanctions, education and reputational pressure.

Thank you !