ETHICAL CONSIDERATIONS ON HATE SPEECH AND AI

HIDA Lecture Series

18. APRIL 2024 I BERT HEINRICHS

JÜLICH
Forschungszentrum
Wandel gestalten

Mitglied der Helmholtz-Gemeinschaft

HATE SPEECH

https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence

WWW.COE.INT HUMAN RIGHTS DEMOCRACY RULE OF LAW ABOUT US English Connect

COUNCIL OF EUROPE

European Commission against Racism and Intolera

Home News About Country Monitoring ECRI Standards Themes Events Resources

You are here: Democracy and Human Dignity › European Commission against Racism and Intolerance (ECRI) › Themes › Hate speech and violence

Hate speech and violence

Hate speech covers many forms of expressions which advocate, incite, promote or justify hatred, violence and discrimination against a person or group of persons for a variety of reasons.

It poses grave dangers for the cohesion of a democratic society, the protection of human rights and the rule of law. If left unaddressed, **it can lead to acts of violence and conflict on a wider scale**. In this sense hate speech is an extreme form of intolerance which contributes to hate crime.

Aware of the **dangerous link between hate speech and violence**, ECRI has always considered that criminal prohibition is necessary when hate speech publicly incites violence against individuals or groups of people. At the same time criminal sanctions should be used as a measure of last resort and, all along, **a balance must be kept between fighting hate speech on the one hand, and safeguarding freedom of speech on the other**. Any restrictions on hate speech should not be misused to silence minorities and to suppress criticism of official policies, political opposition or religious beliefs.

In many instances, ECRI has found that an effective approach to tackling hate speech, in particular cyberhate, is self-regulation by public and private institutions, media and the Internet industry, such as the adoption of codes of conduct accompanied by sanctions for non-compliance. **Education and counter-speech are also equally important** in fighting the misconceptions and misinformation that form the basis of hate speech. Therefore, ECRI considers that effective action against the use of hate speech requires raising public awareness of the importance of respecting pluralism and of the dangers posed by hate speech.

Underreporting of hate speech and hate-motivated violence is another unfortunate feature of these two phenomena. Victims rarely report incidents to the authorities for fear of retaliation or of not being taken seriously, or because they have no confidence in the justice system. This contributes to lack data which makes it difficult to quantify the extent of the problem and take effective measures to address it. **ECRI recommends states to provide practical support to those targeted by hate speech and violence**: they should be made aware of their rights to redress through administrative, civil and criminal proceedings and encouraged to report to the authorities, and receive legal and psychological assistance.

ECRI European Commission against Racism and Intolerance

Themes

- Equality bodies
- Civil society
- Integration and inclusion
- Hate speech and violence
- Legislation against racism and intolerance
- Sexual orientation, gender identity and sex characteristics

@ECRI_CoE

„**Hate speech** covers many forms of expressions which advocate, incite, promote or justify hatred, violence and discrimination against a person or group of persons for a variety of reasons.

It poses grave dangers for the cohesion of a democratic society, the protection of human rights and the rule of law. If left unaddressed, **it can lead to acts of violence and conflict on a wider scale**. In this sense hate speech is an extreme form of intolerance which contributes to hate crime.“

HATE SPEECH

https://hateaid.org/hatespeech/

„Hate speech is a generic term for **group-related misanthropy** expressed verbally or in writing. This includes **racism, sexism and anti-Semitism**. Hate speech includes insults, calls for violence, threats and other statements, **regardless of whether they are punishable by law or not**. However, the term hate speech is not clearly defined and is sometimes also used to refer to general hate on the internet. “

Hate Aid

Das sind wir Wir unterstützen Wir handeln

ETHICAL PRINCIPLES

Beauchamp / Childress [8]2019

Autonomy

Nonmaleficence

Beneficence

Justice

ETHICAL PRINCIPLES

Beauchamp / Childress [8]2019

Autonomy

Nonmaleficence

Beneficence

Justice

ETHICAL PRINCIPLES

Beauchamp / Childress [8]2019

Autonomy

Nonmaleficence

Freedome of Speech

Integrity of the person

SOURCES OF HATE SPEECH

Social Media

Social media can encourage the radicalization of opinions and lead to rapid and often uncontrolled dissemination.

The source of hate speech is always a person who is responsible for their actions.

Possible countermeasures include monitoring social media, deleting posts and sanctioning authors and platform operators (see Network Enforcement Act).

There is a risk of censorship.

Large Language Models

Large language models generate text on the basis of very large amounts of data and can reproduce the distortions contained therein.

The origin of hate speech is an algorithm that generates text in an untraceable way.

Possible countermeasures include the monitoring of LLMs, their manual correction and the sanctioning of providers and users (see AI Act).

There is a risk of innovation being hindered.

SOURCES OF HATE SPEECH

Social Media

Social media can encourage the radicalization of opinions and lead to rapid and often uncontrolled dissemination.

The source of hate speech is always a person who is responsible for their actions.

Possible countermeasures include monitoring social media, deleting posts and sanctioning authors and platform operators (see Network Enforcement Act).

There is a risk of censorship.

Large Language Models

Large language models generate text on the basis of very large amounts of data and can reproduce the distortions contained therein.

The origin of hate speech is an algorithm that generates text in an untraceable way.

Possible countermeasures include the monitoring of LLMs, their manual correction and the sanctioning of providers and users (see AI Act).

There is a risk of innovation being hindered.

AI-BASED APPROACHES FOR SOCAIL MEDIA

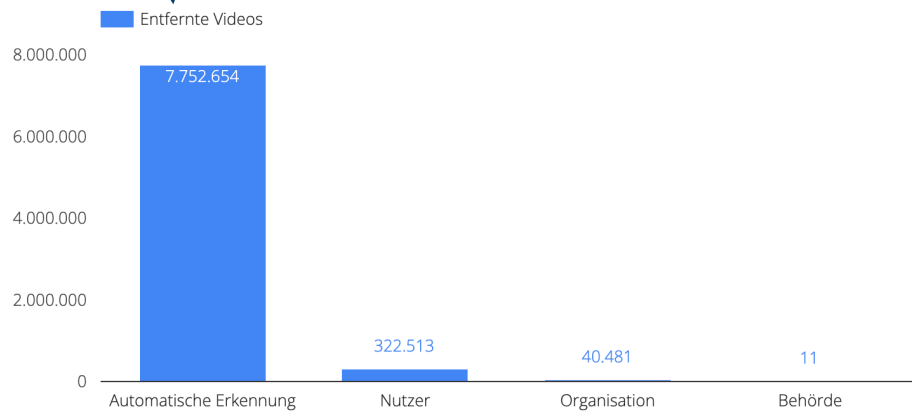
https://transparencyreport.google.com/

YouTube

Automatically removed > 95 %

Entfernte Videos nach Art der ersten Erkennung

This chart shows the volume of videos removed by YouTube, by source of first detection (automated flagging or human detection). Flags from human detection can come from a user or a member of YouTube's Priority Flagger program. Priority Flagger program members include NGOs and government agencies that are particularly effective at notifying YouTube of content that violates our Community Guidelines.

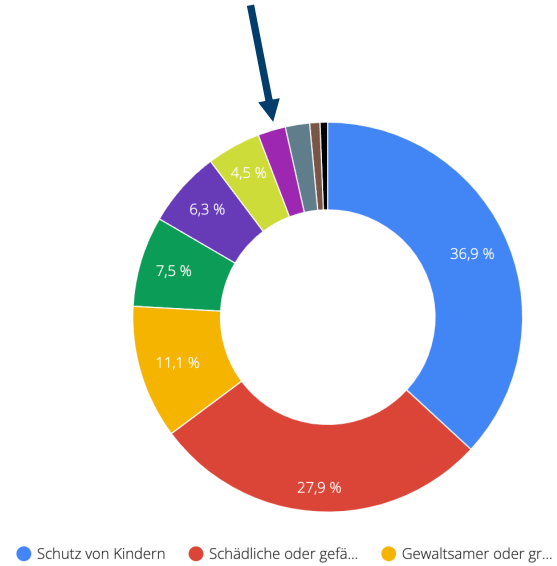


Juli 2023 bis Sept. 2023 | Automatische Erkennung einschließen

Hate speech and offensive content 2,3%

Entfernte Videos nach Entfernungsgrund

Diese Grafik zeigt die Anzahl der von YouTube entfernten Videos nach Entfernungsgrund. Die Entfernungsgründe entsprechen den Community-Richtlinien von YouTube. Bei der Prüfung der gemeldeten Videos werden alle Community-Richtlinien und weitere Richtlinien zugrunde gelegt, unabhängig vom ursprünglichen Beschwerdegrund.



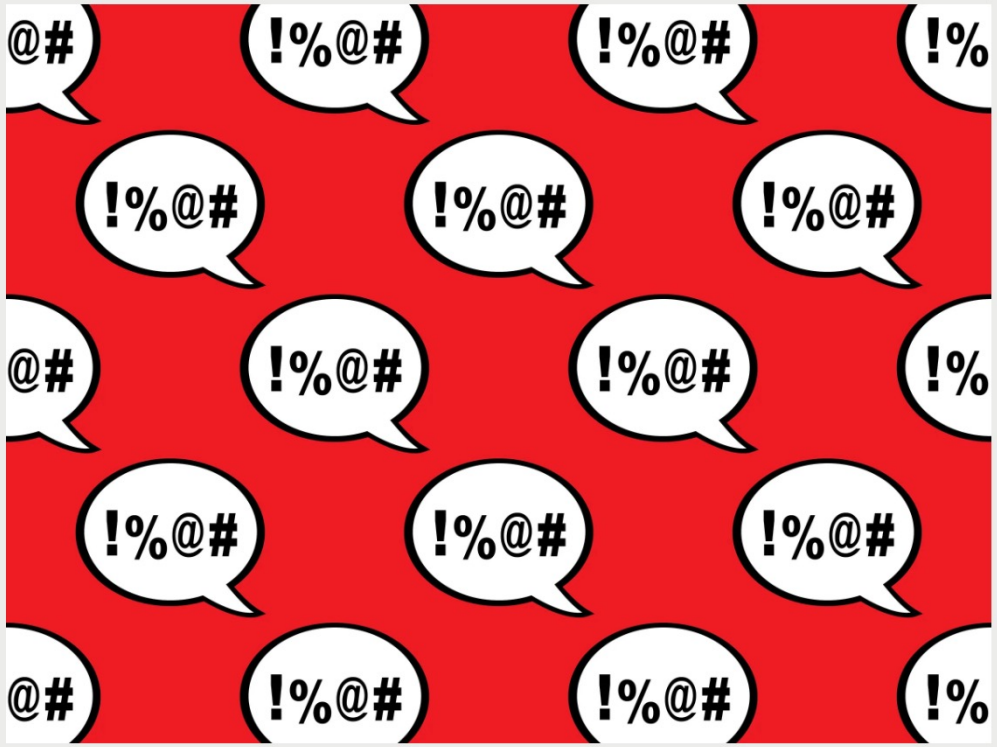
AI-BASED APPROACHES FOR SOCAIL MEDIA

https://spectrum.ieee.org/ai-versus-online-hate-speech

NEWS | ARTIFICIAL INTELLIGENCE

Combating Hate Speech Online With AI › It analyzes the context of social media posts to accurately detect hate speech

BY MICHELLE HAMPSON | 21 FEB 2023 | 2 MIN READ |



Hate speech on social media is increasing, deterring some people from participating while creating a toxic environment for those who remain online. Many different AI models have been developed to detect hate speech in social media posts, but it has remained challenging to develop ones that are computationally efficient and are able to account for the context of the post—that is, determine whether the post truly contains hate speech or not.

A group of researchers in the United Kingdom has developed a new AI model, called BiCapsHate, that overcomes both of these challenges. They describe it in a study published 19 January in *IEEE Transactions on Computational Social Systems*.

AI-BASED APPROACHES FOR SOCAIL MEDIA

https://www.economist.com/the-economist-explains/2023/04/20/why-winnie-the-pooh-makes-xi-jinping-uncomfortable

The Economist

Menu | Weekly edition | The world in brief | Search

The Economist explains

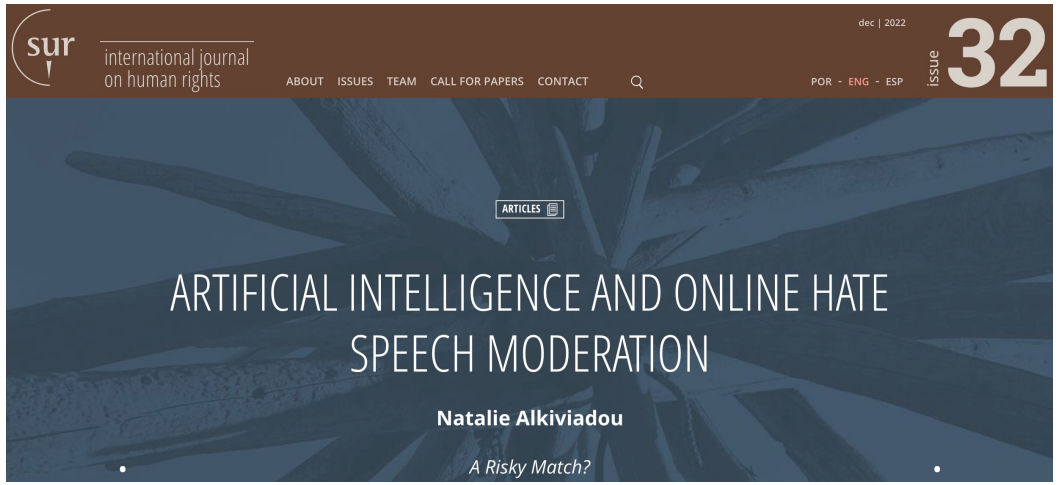
Why Winnie-the-Pooh makes Xi Jinping uncomfortable

China's censors have declared the cuddly bear ursus-non-gratus

IMAGE: ALLSTAR/DISNEY

AI-BASED APPROACHES FOR SOCAIL MEDIA

https://sur.conectas.org/en/artificial-intelligence-and-online-hate-speech-moderation/



„First and foremost, it must be underlined that, as noted by Llanso,52 the above issues cannot be tackled with more sophisticated AI. Moreover, as noted by Perel and Elink-Koren, ‚the process of translating legal mandates into code inevitably embodies particular choices as to how the law is interpreted, which may be affected by a variety of extrajudicial considerations, including the conscious and unconscious professional assumptions of program developers, as well as various private business incentives.‘ Whilst automated mechanisms can assist human moderators by picking up on potentially hateful speech, they should not be solely responsible for removing hate speech. Biased training data sets, the lack of relevant data and the lack of conceptualization of context and nuance can lead to wrong decisions, which can have dire effects on the ability of minority groups to function equally in the online sphere.“

SOURCES OF HATE SPEECH

Social Medien

Social media can encourage the radicalization of opinions and lead to rapid and often uncontrolled dissemination.

The source of hate speech is always a person who is responsible for their actions.

Possible countermeasures include monitoring social media, deleting posts and sanctioning authors and platform operators (see Network Enforcement Act).

There is a risk of censorship.

Large Language Models

Large language models generate text on the basis of very large amounts of data and can reproduce the distortions contained therein.

The origin of hate speech is an algorithm that generates text in an untraceable way.

Possible countermeasures include the monitoring of LLMs, their manual correction and the sanctioning of providers and users (see AI Act).

There is a risk of innovation being hindered.

AI-BASED APPROACHES FOR LLMS

https://time.com/6247678/openai-chatgpt-kenya-workers/

TIME

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image

„The premise was simple: feed an AI with labeled examples of violence, hate speech, and sexual abuse, and that tool could learn to detect those forms of toxicity in the wild. That detector would be built into ChatGPT to check whether it was echoing the toxicity of its training data, and filter it out before it ever reached the user. It could also help scrub toxic text from the training datasets of future AI models.

To get those labels, OpenAI sent tens of thousands of snippets of text to an outsourcing firm in Kenya, beginning in November 2021. Much of that text appeared to have been pulled from the darkest recesses of the internet.“

AI-BASED APPROACHES FOR LLMS

https://techcrunch.com/2023/04/12/researchers-discover-a-way-to-make-chatgpt-consistently-toxic/

- Manual "cleansing" cannot prevent LLMs from being used as tools for immoral purposes.
- In addition to the toxic distortions that are contained in the training data and that have to be laboriously removed manually, there is a toxicity of the application that can hardly be prevented.
- Powerful tools can cause great damage.

AI

Researchers discover a way to make ChatGPT consistently toxic

TechCrunch

Kyle Wiggers @kyle_l_wiggers / 3:00 PM GMT+2 • April 12, 2023

Image Credits: STEFANI REYNOLDS/AFP / Getty Images

OUTLOOK

- Freedom of speech is a central ethical principle and enjoys a high level of protection in democratic states.
- The integrity of the person is an equally central ethical principle and enjoys a similarly high level of protection.
- In case of conflicting principles, well-founded trade-offs must be found, which often requires case-by-case assessments.
- Technological solutions can help to detect problematic cases, but they cannot weigh up ethical considerations.
- Fundamental rights are linked to the ability to take responsibility for one's own actions.
- Anyone who uses violence against others must be held accountable.
- Whether a statement is still covered by freedom of expression or should count as an act of violence is open to debate.
- Technology-based solutions are needed for support. However, full automatization is problematic.
- The fight against hate speech must be embedded in programs for strengthening democracy and media literacy.

Thank you for your attention!

Bert Heinrichs

b.heinrichs@fz-juelich.de

Der Mensch als Industriepalast (Detail), Fritz Kahn 1926, Franckh-Kosmos

Forschungszentrum Jülich
Institut für Neurowissenschaften und Medizin
Gehirn und Verhalten (INM-7)

Arbeitsgruppe
Neuroethik und Ethik der KI

Neuroethics and Ethics of AI
Working Group

Brain and Behaviour (INM-7)
Institute of Neuroscience and Medicine
Research Center Jülich