Artificial Intelligence to Counter Cyber-Terrorism

Serena BIANCHI[1], Marina MANCUSO[2], Caterina PATERNOSTER[3], George KALPAKIS[4], Theodora TSIKRIKA[5], Stefanos VROCHIDIS[6], Denitsa KOZHUHAROVA[7], Bernhard JAEGER[8]

[1] Research Department, SYNYO GmbH, Vienna, Austria
serena.bianchi@hotmail.it
[2] Transcrime Università Cattolica del Sacro Cuore, Milan, Italy
marina.mancuso@unicatt.it
[3] Transcrime Università Cattolica del Sacro Cuore, Milan, Italy
caterina.paternoster@unicatt.it
[4] Information Technologies Institute, Center for Research and Technology Hellas, Thermi-Thessaloniki, Greece
kalpakis@iti.gr
[5] Information Technologies Institute, Center for Research and Technology Hellas, Thermi-Thessaloniki, Greece
theodora.tsikrika@iti.gr
[6] Information Technologies Institute, Center for Research and Technology Hellas, Thermi-Thessaloniki, Greece
stefanos@iti.gr
[7] Law and Internet Foundation, Sofia, Bulgaria
denitsa.kozhuharova@netlaw.bg
[8] Research Department, SYNYO GmbH, Vienna, Austria
bernhard.jaeger@synyo.com

Abstract

This paper discusses the role of disruptive and innovative technologies for countering the spread of terrorist online content (TCO). In particular, it focuses on the use of Artificial Intelligence (AI) in support to Host Service Providers (HSPs) and Law and Enforcement Agencies (LEAs). The violent and terrorist content is more and more disseminated online taking advantages of the opportunities offered by Internet. The diffusion of terrorist propaganda has a negative impact on the civil society and poses several risks. For this reason, the European institutions published in 2021 the Regulation (EU) 2021/784 to address the misuse of hosting services for the dissemination to the public of TOC. It regulates the measures to be applied by HSPs and Member States' authorities in order to identify and ensure the quick TOC removal and to facilitate cooperation with each other and Europol. In order to be compliant with these dispositions, AI-based disruptive technologies can provide LEAs and HSPs, especially the small and micro-ones, a concreate support. The implementation of the Regulation and the use of AI technologies have legal and ethical implications that have to be considered. The paper is based on the work and preliminary research conducted in the framework of the European funded project ALLIES, "AI based framework for supporting micro and small Hosting Service Providers (HSPs) on the report and removal of online terrorist content", Grant Number 101080090.

Index terms: Artificial Intelligence, Counter Extremism, Cyberterrorism, Ethical and Legal Framework, Online Radicalisation

1. Introduction

Online media channels deeply changed the way how people communicate, work, interact and live. Together with innovation and unprecedented potentialities, they also introduced new threats into our society, changing the modus operandi and the structures of terrorist organisations.

As highlighted by the European Parliament (2021), the dissemination of terrorist content is one of the most widespread and most dangerous forms of misuse of online services in the field of internal security [1]. Online and social media channels have been broadly used in the past years by terrorist organisations in order to spread violent content, to train, recruit and motivate individuals, and last but not least, to finance terrorist organisations [2, 3, 4].

Similarly as for "terrorism", the definition of "cyber-terrorism" is still very controversial. Some authors define cyber-terrorism as the premeditated use of disruptive activities, or the threat thereof, against computers and/or networks, with the intention to cause harm or further social, ideological, religious, political or similar objectives, or to intimidate any person in furtherance of such objectives. [5].

In this paper, the authors will consider cyber-terrorism in broader terms, including also the capacity of carrying out cyber-related crimes, such as the **spread of online violent content**, training and recruiting of individuals and terrorist financing (following the Directive (EU) 2017/541) [6].

In particular, the focus is given to the spread of online terrorist generated content, having as scope one or more of the following actions (according to Art.2(7) of the Regulation (EU) 784/2021):

a) inciting or advocating, including by glorifying, the commission of terrorist offences, thereby causing a danger that such acts be committed.
b) encouraging the contribution to terrorist offences.
c) promoting the activities of a terrorist group, in particular by encouraging the participation in or support to a terrorist group within the meaning of Article 2(3) of Directive (EU) 2017/541.
d) instructing on methods or techniques for the purpose of committing terrorist offences.

The paper is structured in four main sections. In the first one, the authors discuss the threat posed by terrorism online content (TCO) mainly for the civil society. In the second one, the TCO EU Regulation is presented in order to understand what are the main new dispositions foreseen at EU level for preventing and fighting against TCO. After that, and considering this initial framework, in Sections 3 and 4 the authors analyse two main components related to TCO: (a) the technical capabilities that innovative technologies can provide as support to the Law Enforcement Agencies (LEAs), as well to private service providers, and (b) the ethical and legal component to be considered when dealing with these topics.

2. The Threat posed by Terrorist Online Generated Content

Before diving into how innovative technologies can support the detection and removal of violent and extremist content online, it is of importance to highlight why terrorist online generated content poses such a big threat to our civil society and what are the risks related to it.

On the one hand, the establishment and vast evolution of new communication channels via the Internet opened new possibilities for the overall distribution of terrorist-generated online content. While there was previously almost no alternative to dissemination via text, radio or television, the Internet offered countless possibilities for dissemination. More specifically, livestreaming, pictures and videos are widely used as a format of disseminating terrorist-generated content. Several extremists and members of terrorist organisations adapted to the new communication channels and, starting already from the early 2000s, numerous photos and videos were distributed praising the hijackers and attackers of terrorist attacks - see the most recent attacks in Christchurch; Buffalo; or

the synagogue shooting in Halle, Germany, along with another attack on a synagogue in Poway, California; a racist attack at a Walmart in El Paso, Texas; and at a mosque in Bærum, Norway; or again the several violent videos shared by ISIS members starting from 2014.

On the other hand, the online world has opened up new opportunities for radicalisation, even for those who lack physical connections to radicalised individuals or environments, increasing the chance of self-radicalisation [7, 8]. Social media can provide an outlet for vulnerabilities that stem from offline sources, helping to compensate them. For example, online spaces are able to draw individuals who are feeling socially alienated or isolated, who see in the online world an alternative social setting where they can express their frustrations. In an environment where an abundance of information and propaganda is present, connecting with people who share similar extreme views can reinforce radical thinking by providing validation, and also legitimising the use of violence [9, 10, 11].

The spread of violent content online therefore implies the **facilitation** of **inciting terrorist activities** and **glorifying** such proceedings, by distributing various formats, such as photos, videos or texts, and consequentially, promoting subtle **ideological indoctrination** to compel others to commit terrorist acts [1]. Similarly to the commercial brand e-communication strategies, also terrorist organisations and ideologies make use of marketing tools to spread their ideas more effectively. This phenomenon is enabled and facilitated by creating the highest possible reach of a particular message through the social media channels and websites. That way the range of indirect online recruitment, which is clearly expressed through indoctrination, is increased and the pool of audience drastically enlarged. Besides social media channels extremists increasingly utilised another powerful tool for their activities during the last decade. Gaming and related platforms have grown to become some of the world's largest entertainment industries, providing extremist groups with significant opportunities for recruitment and organisation [12]. The video game Salil alSawarem (The Clanging of Swords) can be considered for instance as modern example of high-quality propaganda. It is a "first-person shooter" game that imitated the popular Grand Theft Auto franchise and was designed to gain attention for ISIS. The game's trailers were released on numerous websites and platforms, including YouTube, which had 3.5 billion views per month on gaming channels alone when the game was released in 2014. The terrorists' use of cinematic productions, social media, and the appeal of a videogame demonstrates their tactic of exploiting popular culture to make their propaganda go viral and reach their target audience [13]. Also, far-right extremists are increasingly present in online gaming, while the industry's lack of content moderation, hidden metrics, and avoidance of the issue are impeding efforts to assess and combat the problem [12]. Extremist groups are spreading abusive messages and forming relationships in games ranging from military shooters like Call of Duty to open creative environments like Roblox. According to a 2019 report by the Anti-Defamation League, 23%, reported exposure to discussions of white supremacist ideology [14].

In the last years, the amount, and consequently the detection, of TCO has sharply increased. Between July 2015 and 2018, the EU Internet Referral Unit[1] (EU IRU) received over 50,000 decisions of referrals to service provides about terrorist content on their platforms. In 2021, Twitter reported over 1.8 million accounts for violations of their terms of service in relation to the promotion of violence and extremism [15], while in 2022 Facebook removed more than 56 million pieces of content containing terrorist propaganda [16].

In this content, rapid actions using disruptive and innovative technologies become fundamental means for countering the spread of TCO.

[1] Europol established the EU IRU in 2015 to actively scan the Internet for terrorist content and then refer it to the hosting service providers, accompanied by an assessment.

3. TCO Regulation & the role of Host Service Providers (HSPs)

The European Commission carried out different consultations as part of the impact assessment SWD (2018) 409 [17], to understand the stakeholders' view on TCO matter. Results show that HSPs are mostly and noticeably damaged by the dissemination of TCO. This affects not only their reputation, but also the relations with their users, and all the relevant stakeholders, who support them in their business – among others, also payment processors.

In this continuously evolving context, where the spread of TCO is an actual emerging threat, negatively impacting on several HSPs, including micro, medium, and small enterprises, what is the role of the hosting service providers (HSPs) and who is responsible for the identification, detection and removal of TCO?

The European Commission implemented several measures for tackling the dissemination of TCO, such as the EU Internet Forum, Radicalisation Awareness Network, the European, Strategic Communications Network, and most importantly, the above-mentioned EU Internet Referral Unit (IRU).

However, after noticing that *referrals alone will not be able to achieve the necessary impact – particularly with regards [sic] to volume and speed of response*, the Commission proposed the TCO (Terrorist Content Online) Regulation, which envisages the transformation of extant co-regulation and voluntary self-policing by HSPs **into a framework of obligations** (to respond to referrals; to comply with removal orders; to implement proactive measures) bolstered by sit-up-and-take-notice sanctions. [18]

As highlighted by the EU-funded project ALLIES, the fact that these new requirements set out one hour timeframe for reaction leads to the need for a prompt adaptation and implementation of measures by the HSPs. Additionally, any HSP exposed to terrorist content as per Art. 5, para 4 of the TCO Regulation should implement the so-called specific measures, making the balance between avoiding fines and respecting the legal principles a very complicated and demanding matter for the HSPs. Also, as outlined in the points above, with the new Regulation alongside hopes for better results in the removal of TOC, significant challenges likewise arise for HSPs in meeting the new obligations and expectations. The adaptation for micro & small HSPs will be far more challenging than for larger HSPs, due to the limited capacity of capital and human resources. In this regard, in the next paragraph, the authors will analyse how Artificial Intelligence can practically support the HSPs in identifying, tackling, and removing online terrorist content.

Additionally, it should be considered, that even before the Regulation entered into force, 61 human rights organisations have officially stated their position against the new piece of legislation due to their concerns about how it will affect the freedom of expression, the rights to information and privacy, and the rule of law in general.[2] Similar concerns have been outlined by the European Data Protection Supervisor in 2019 that was mainly focused on the need for safeguards and prevention measures for conducting profiling of any kind, thus recommending human oversight and verification mechanisms as the best options.[3] These questions, together with the delicate topics on the potential misuse of AI tools applied for the removal of TOC will be analysed in paragraph 4.

4. Artificial Intelligence for countering cyber-terrorism

Given the extensive subversive use of the Internet and the various underlying available online channels by terrorist and extremist groups for the dissemination and propagation of terrorism-related

[2] Joint letter of 61 human rights organisations addressed to the European Parliament, (2021) accessed 03.01.2022.

[3] European Data Protection Supervisor, Formal comments of the EDPS on the Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online (2019).

multilingual and multimodal content using mainly the capabilities offered by HSPs, the adoption of AI-based disruptive technologies by hosting providers and law enforcement has become ever more significant to counter such threats. On the one hand, modern LEAs can leverage the available innovative solutions to collect open-source intelligence and process large amounts of data rapidly, hence, to analyse online content relevant to an ongoing investigation in a timely manner, whilst also allocating the available resources to cover additional operational needs. In this context, the LEAs are equipped with additional means to detect and identify subversive content online, including material that poses an imminent threat to life, and request its removal from the respective hosting providers under the TCO Regulation.

On the other hand, the HSPs can benefit by enhancing their arsenal with tools permitting the continuous automatic monitoring and analysis of the big amounts of data uploaded by the numerous users of their platforms, resulting in the detection and removal of such content in a more efficient manner, requiring fewer human resources. This is of particular interest for the ALLIES project, which focuses on HSPs of small and medium size characterised by limited capacity to respond to the abuse of their online spaces. In this context, disruptive technologies can also contribute to the capacity of HSPs to respond to removal orders issued by law enforcement within the timeframe set by the one-hour rule; AI-based tools can be used to automatically analyse removal orders acting as a decision-support mechanism facilitated by the HSP content moderators.

Several domains of AI including (but not limited to) deep neural networks and machine learning based on supervised or unsupervised learning approaches, can facilitate the fight towards countering the multimodal terrorist content online. Natural Language Processing (NLP) solutions provide useful insights by analysing the multilingual online textual content and extracting, disambiguating, linking, and semantically enhancing concepts and named entities [19] associated with terrorism- and extremism-related content. Automatic speech recognition tools process audio and visual files to produce accurate transcriptions for languages of interest that can be subsequently analysed by NLP technologies [20]. Computer vision models employing deep learning algorithms analyse videos and images to detect and recognise objects [21], concepts, human behaviour and activities [22] relevant to the domain of terrorism and extremism.

The indicators produced by the separate analysis of the textual, audio, and visual modality is leveraged by multimodal classifiers that automatically categorise the online data into a set of predefined categories relevant to the terrorism and extremism domain [23]. Powerful explainable AI techniques [24] leveraging the outputs produced by the aforementioned multimodal AI models support human interpretation, providing HSP moderators and LEA analysts with reasoning and explanation on the results delivered by the AI solutions, thus helping towards making informed decisions related to the removal of abusive content from the online space. Finally, threat assessment tools [25] help HSP moderators and LEA analysts towards assessing the severity of threats posted online, thus enforcing the appropriate measures including the removal of such material.

5. The Legal and Ethical Framework

Based on the above-mentioned considerations, it should be also underlined, that regulating the digital realm can oftentimes lead to interference with fundamental rights and presents a difficult conundrum of striking the balance between fundamental rights respect and guaranteeing public order and security. This is applicable also to the case of the TCO Regulation practical implementation. The latter outlines a number of obligations to HSPs leaving them with a little margin for independent decision-making, most notably in the case of removal order receipt. According to Art. 3 para 3 of the TCO Regulation, HSPs need to act promptly upon a removal order receipt, taking down the respective content in an hour, or in case of removal order issued by a competent authority from another EU Member State – 72 hours (Art. 4, para 3, TCO Regulation). In this case HSPs are not provided with

a discretion whether or not to comply with the removal order, the Regulation provides quite limited circumstances in which removal could be refused – if the removal order itself is erroneous or does not provide sufficient information (Art. 3, para 8, TCO Regulation), where sufficient information is defined as "*location of that content, by indicating the exact URL*" [26].

At the same time, the new TCO Regulation obliges HSPs to establish a complaint mechanism where content providers could contest the removal of their content or the disabling the access to it (Art. 10 para 1, TCO Regulation). This being said, it should be noted that any specific action a HSP would take in connection to applying measures countering the misuse of the services it provides is intrinsically linked to receiving removal orders from the competent authorities, as per Art. 5 of the TCO Regulation. Thus, the situation where HSPs are put in requires them to deal with complaints and objection to their actions, while the latter are being prompted by the assessment made by the competent authorities. The Regulation also establishes in Art. 9 the availability of legal remedies for both HSPs and content providers challenging a removal order before the competent courts in the respective EU Member State. It could be argued that going to court for contesting a removal order on behalf of a HPS or a content provider might render the rights holders unwilling due to lengthy and costly procedures, still it remains to be seen how often and to what extent the available legal remedy would be used in practice.

Considering the fundamental rights and the practical implementation of the TCO Regulation, a few remarks need to be made with respect to the enforcement of the right to freedom of expression. The text of the Regulation explicitly states that it should be applied in a sense that changes the postulates of EU law with respect to freedom of expression and information [26]. This is further elaborated in the preamble of the Regulation, where it is noted that "*expression of radical, polemic or controversial views in the public debate on sensitive political questions*" can never be qualified as terrorist content. What is more, the preamble calls for due consideration of the freedom of expression, alongside the right to information, the freedom and pluralism of media and the freedom of arts and sciences, whenever an assessment is being made whether certain types of content should be classified as terrorist or not.

Such an assessment becomes more complex once AI-powered tools are involved in the flagging of terrorist content online, which is the case of the ALLIES project striving the equip micro and small HSPs with the necessary knowledge, skills, and tools to ensure the practical implementation of the TCO Regulation. One of the risks that the ALLIES project team would address is how to prevent the planned tools of bias and discrimination, which may hamper freedom of expression. The latter could occur in case information leading to discrimination is being processed by the AI algorithm and contributes to the determination of whether a certain piece of content is in fact terrorism-related [27]. Which type of information could lead to unfavourable treatment is prescribed by EU law (Directive 2000/43/EC and Directive 2000/78/EC), namely information of:

- racial or ethnic origin,
- religion or belief,
- disability,
- age,
- sexual orientation.

An example of such a case would be if a video is removed from a video hosting platform due to the fact that the AI tool has labelled it as terrorism-related solely because the author of the video is of particular ethnic origin. However, the workings of an AI tool are not that simple, and one may fall victim of discrimination and unfair treatment due to "*profiling identities based on a combination of behavioural and demographic characteristics*" [27].

The application of AI in the field of tackling TOC is not that gloomy despite the risks outlined above. Although at the moment there is no legal framework at EU level regulating the use and application of AI per se, there is robust ethical guidance available postulating best practice and golden

standards which, when followed, provide high guarantees to the rights and freedoms of the individuals. One of the main directions EU policy takes is that of trustworthy AI [28]. The latter is achieved through the mainstreaming of the following key requirements [28]:

- human agency and oversight,
- technical robustness and safety,
- privacy and data governance,
- transparency,
- diversity, non-discrimination and fairness,
- societal and environmental wellbeing, and
- accountability

Aiming to minimise the risk of bias and discrimination in the case of the AI-supported TCO Regulation implementation, the key requirements presented above should be transformed into practical parameters. For example, the requirement of 'diversity, non-discrimination and fairness' might be translated into a practice where robust, diverse, and representative data sets are used for the training and testing of AI algorithms [28]. This could be further enhanced by putting into practice the requirements 'human agency and oversight' where personnel of diverse backgrounds are involved in the training, testing and application of the AI tools [29]. The latter requirement is of high importance in dealing with terrorism content online, as identifying potential criminal behaviour is a decision-making process that cannot be entirely allocated to machines considering the risk of potential harm. To this end, it is recommended that the human-in-command approach is utilised [29], meaning that a person of appropriate knowledge and experience is put in a position to supervise, closely monitor and decide how the results yielded by the AI algorithm should be interpreted, and what kind of actions should follow-up. Additional oversight mechanisms in place could further the trustworthiness of the AI system and serve as an additional guarantee against bias and discrimination occurring [29].

6. Conclusion

Based on the initial findings of the EU-funded project ALLIES, this paper described the strengths and weaknesses of disruptive technologies, with particular focus on AI used by LEAs and HSPs for countering cyber-terrorism.

Technologies can, on the one side, contribute to the capacity of HSPs to respond to removal orders issued by law enforcement within the timeframe set by the one-hour rule (following the TCO Regulation). Thus, HSPs can benefit by enhancing their arsenal with tools permitting the continuous automatic monitoring and analysis of the big amounts of data uploaded by the numerous users of their platforms, resulting in the detection and removal of such content in a more efficient manner, requiring fewer human resources. On the other side, they allow to equip LEAs with additional means to detect and identify subversive content online and request its removal from the respective hosting providers under the TCO Regulation. Overall, AI models support by providing HSP moderators and LEA analysts with reasoning and explanation on the results delivered by the AI solutions, thus helping towards making informed decisions related to the removal of abusive content from the online space. Finally, threat assessment tools help HSP moderators and LEA analysts towards assessing the severity of threats posted online, thus enforcing the appropriate measures including the removal of such material.

However, the support provided from innovative technologies, should be properly framed in a well-grounded legal framework, which allow to avoid bias and discrimination. The authors provide a way forward in order to strengthen this issue, by putting into practice the requirements 'human agency and oversight' where personnel of diverse backgrounds are involved in the training, testing and application of the AI tools.

References

[1] K. Luyten, "Addressing the dissemination of terrorist content online," European Parliamentary Research Agency, 2021.

[2] M. Mengu and S. Mengu, "Violence and Social Media," Athens Journal of Mass Media and Communications, vol. 1, no. 3, pp. 211-228, 2015.

[3] A. Tsesis, "Social Media Accountability for Terrorist Propaganda," Fordham Law Review, vol. 86, no. 2, pp. 605-632, 2017.

[4] T. Keatinge and F. Keen, "Social Media and (Counter) Terrorist Finance: A Fund-Raising and Disruption Tool," Studies in Conflict & Terrorism, vol. 42, no. 1-2, pp. 178-205, 2019.

[5] F. D. University., "Cybersecurity and Cyber Terrorism," [Online]. Available: https://online.fdu.edu/program-resources/cybersecurity-and-cyber-terrorism/. [Accessed 2023 April 2023].

[6] EUROPEAN PARLIAMENT AND COUNCIL, "eur-lex.europa.eu/," 17 March 2017. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017L0541&rid=6. [Accessed 15 March 2023].

[7] A. Meleagrou-Hitchens and N. Kaderbhai, "Research Perspectives on Online Radicalization. A Literature Review, 2006-2016," International Centre for the Study of Radicalization (ICSR), London, 2017.

[8] E. Charlie and L. Gribbon, "Pathways to Violent Extremism in the Digital Era," The RUSI Journal, vol. 158, no. 5, pp. 40-47, 2013.

[9] G. N. Mølmen and J. A. Ravndal, "Mechanisms of online radicalisation: how the internet affects the radicalisation of extreme-right lone actor terrorists," Behavioral Sciences of Terrorism and Political Aggression, pp. 1-25, 2021.

[10] L. S. Neo, "An Internet-Mediated Pathway for Online Radicalisation: RECRO," in Combating Violent Extremism and Radicalization in the Digital Era, Hershey, PA, USA, IGI Global, 2016, pp. 197-224.

[11] P. R. Neumann, "The trouble with radicalization," International Affairs, vol. 89, no. 4, pp. 873-893, 2013.

[12] L. B. Galen Lamphere-Englund, "State of Play on Gaming & Extremism – Reviewing the literature on gaming & extremism," 6 October 2021. [Online]. Available: https://drive.google.com/file/d/1WEq4OjtqZYdltAB0SK46M88gFF863jWs/view. [Accessed 28 April 2023].

[13] A. Al-Rawi, "Video games, terrorism, and ISIS's Jihad 3.0," Terrorism and Political Violence, vol. 30, no. 4, pp. 740-760, 2018.

[14] C. Ingersoll, "Free to Play? Hate, Harassment and Positive Social Experiences in Online Games 2020," Anti-Defamation League, 2020.

[15] Twitter, "Transparency - Rules Enforcement," Twitter, July-December 2021. [Online]. Available: https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec. [Accessed 27 04 2023].

[16] Meta, "Meta - Transparency Center," Meta, 2023. [Online]. Available: https://transparency.fb.com/data/community-standards-enforcement/dangerous-organizations/facebook/. [Accessed 27 04 2023].

[17] European Commission, "https://eur-lex.europa.eu/," 12 September 2018. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=SWD:2018:408:FIN. [Accessed 14 March 2023].

[18] G. Robinson, "The European Commission's Proposal for a Regulation on Preventing the Dissemination of Terrorist Content Online," 31 January 2019. [Online]. Available:

https://eucrim.eu/articles/commission-proposal-regulation-preventing-dissemination-terrorist-content-online/. [Accessed 2023 April 28].

[19] "Autoregressive Structured Prediction with Language Models," AutorarXiv preprint arXiv:2210.14698, 2022.

[20] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," Interspeech, vol. 2013, pp. 2345-2349, 2013.

[21] D. Touska, K. Gkountakos, K. Ioannidis, T. Tsikrika, S. Vrochidis, I. Kompatsiaris, "Graph-Based Data Association in Multiple Object Tracking: A Survey," in In Proceed of the 29th International Conference on Multimedia Modeling, 2023.

[22] K. Gkountakos, D. Touska, K. Ioannidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "Spatio-temporal activity detection and recognition in untrimmed surveillance videos.," in In Proceedings of the 2021 International Conference on Multimedia Retrieval, 2021.

[23] "Domain-aligned Data Augmentation for Low-resource and Imbalanced Text Classification," in Proceedings of the 45th European Conference on Information Retrieval (ECIR'23), Dublin, 2023.

[24] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.Z. Yang, Explainable artificial intelligence. Science robotics, vol. 4, no. 37, 2019.

[25] O. Theodosiadou, D. Chatzakou, T. Tsikrika, S. Vrochidis and I. Kompatsiaris, "Real-time Threat Assessment based on Hidden Markov Models.," 2023.

[26] European Parliament and the Council, Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online (Text with EEA relevance), Official Journal of the European Union.

[27] European Union Agency for Fundamental Rights, Bias in Algorithms – Artificial Intelligence and Discrimination, Vienna: Luxembourg: Publications Office of the European Union, 2022.

[28] European Commission, "WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust," 2020.

[29] High-Level Expert Group on Artificial Intelligence, "ETHICS GUIDELINES FOR TRUSTWORTHY AI," European Commission, Brussels, 2019.

[30] K. Luyten, "Addressing the dissemination of terrorist content online," European Parliamentary Research Service, 2021.

[31] N. Stylianou, D. Chatzakou, T. Tsikrika, S. Vrochidis, I. Kompatsiaris, "Domain-aligned Data Augmentation for Low-resource and Imbalanced Text Classification," in ECIR'23, Dublin, 2023.