Detecting and countering radicalisation online

A review of recent literature

by Michael Winberg

Resumé

Information kan användas för att på olika sätt påverka en aktör (exempelvis en individ) att agera på ett visst sätt. Syftet med denna artikel är därför att skapa en överblick över forskningen kring att upptäcka radikalisering på Internet, särskilt sociala medieplattformar, och hur processen kan förhindras. Sammantaget är delningen av både våldsamt och icke-våldsamt extremistiskt innehåll beroende av att det finns en efterfrågan där flera faktorer samverkar för att radikalisera en person. När en individ på ett moraliskt plan frånkopplas ifrån samhället passeras en tröskel i radikaliseringsprocessen som kan utnyttjas av illvilliga grupperingar. Framtida forskning behöver koncentreras till att förstå hur narrativ formas och levereras, samt hur dessa kan bemötas.

THE STORMING OF the United States Capitol in January and the succeeding blocking of (then serving) U.S. president Donald Trump from various social media platforms has once again highlighted how social media can be used to radicalise citizens and influence them to act in a violent manner. Information has long been an important part of society where it can be used to, by various means, influence an actor (state, company or even an individual) to act in a certain way. In modern times, this has resulted in information being a dimension where war can be waged between various actors, state and non-state, where it is not necessarily the military adversary who is the recipient but instead the state's population.

The spread of information can also, thanks to modern technology, occur immediately using social media, which means that an individual or organisation can disseminate an image, video, or message in text on hundreds of accounts with the help of just one app. The same method may also be used to convince individuals to conduct certain acts such as help spreading propaganda, donating money, conducting terror acts, or even convincing an individual raised in a western democracy to travel abroad to join a terrorist organisation.

Various nations such as United States, United Kingdom, and Sweden have established national centres for studying and countering terrorism, centres for confronting cyber-related threats as well as strengthening the psychological defence. Such centres may share a mutual interest in detecting and countering violent extremist activity online based on their common "battlespace" (i.e. cyberspace). It could be reasoned that a natural response to countering radicalising content on the internet would be to focus on detecting and hindering the actual crime (the terrorist act) and to lesser extent track from where the propaganda content is derived or how the narrative is constructed. Nonetheless, since radicalisation and extremism know no ideological boundaries, it is important to have a broader perspective of radicalisation that occurs online, not just the "traditional" jihadist content. As such there is an argument to be made that there is a need to be capable not only of detecting or restricting content deemed to be harmful, but also countering the narrative within a wider spectrum of extremism.

Countering radicalisation, or extremism, should in this case be understood as limiting the spread of extremist content, both violent and non-violent, and with that the extremist narrative. The literature reviewed consists of 19 articles published between 2013 and 2020. The articles have been chosen based on containing keywords like extremism and radicalisation and the applicability on today's cyber landscape; meaning the migration from internet discussion boards (often referred to as forums) to social media apps, which mainly benefit articles published recently, hence, positively affecting the validity of the research.

The purpose of this review is therefore to go through current literature on detecting radicalisation on the internet, in particular social media platforms, and establish how it may be prevented. Furthermore, gaps will be identified and addressed when suggesting future areas of research. First, we define what radicalisation and extremist content is, second, how radicalisation does occur on internet, third, how radicalising content may be detected online, fourth, what methods can be used to restrict access to the content and finally, how a counter-narrative may be employed.

Defining radicalisation and extremist content

The words "radicalised" and "extremist" are frequently used in news reporting about suspected terrorists being arrested, about the vulnerability of young immigrants in the suburbs or when proposals of new laws that are supposed to "combat" radicalism are presented by governments. Although common words used daily, there is different understanding in what a person tries to say when labelling something "radical" or "extreme".

Baugut and Neumann state that academia overall agrees on radicalisation being a process that gravitates around ideology. According to the authors, researchers distinguish between radicalisation as being "cognitive", "behavioural" or "violent" using various complex models.[1] Braddock goes a bit further and suggests that radicalisation could be seen as a persuasive process where the consistent exposure of an extremist message turns the individual towards adopting attitudes consistent with an ideology.[2] The Internet in fact enables the process in radicalising individuals by providing access to extremist content as well as online groups where they found support.[3] An individual may as such become radicalised in his or her way of thinking and reasoning, how she acts and as a result violently acting out on society. Important to note is that an individual can possess radical thoughts without acting out on these thoughts in a violent manner.

Von Behr, Reding, Edwards & Gribbon draw on the UK Government definition which defines radicalisation as "the process by which a person comes to support terrorism and forms of extremism leading to terrorism" and "extremism as vocal or active opposition to fundamental [...] values".[4] Although supporting the notion of radicalisation being a process, the UK Government definition focuses on the individual acting out her beliefs, not accounting for the individual and her thoughts, unlike the researchers. Hence, radicalisation is a process where an individual transforms into an extremist, adapting an ideological belief system which may manifest itself in various acts, including violent. Along these lines "online radicalisation" is the radicalisation process occurring within the cyber domain.

In this review the terms "radical" and "extremist" content will be used interchangeably since the content serves the same purpose: convincing and radicalising the individual. The word "violent" denotes content that depicts any type of violence aimed against a human being. It may or may not be with the purpose of radicalisation of an individual.

Radicalisation and the internet

As previously mentioned, it is not only state actors who have the capability to spread well-produced content on various internet platforms. Throughout the world, various extremist groups have created state-of-the-art media wings that flood the internet with professionally edited videos, magazines, and music with the goal to influence, radicalize and recruit new supporters. The public also seems to have become more inclined to document, share images and video from terrorist acts through social media which in turn have led to amplifying the media coverage for single terrorists as well as extremist groups. This meaning that the state, society, and the private sector all together have an impact on the way radicalising content can travel through internet and influence people.[5] For instance, the time between a terrorist event and the act being disseminated online seems to be shrinking rapidly, which in turn has resulted in authorities and social media platforms trying to intervene with varied results.

In 2013 the Somali terrorist organisation al-Shabaab used twitter to publish updates during an ongoing attack--not just to deliver news to its followers, but also to control the narrative about the attack from the beginning.[6] In France 2016, a radicalised male posted on Facebook how he killed a police officer and the officer's wife. In Christchurch 2019, the attacker broadcasted the attack live on Facebook without the stream being stopped, according to the company because no one reported the stream to moderators.[7] It is theorised that a terrorist act, and the spreading of imagery from it, may trigger other individuals to commit own acts of terror, even in other countries. This was especially highlighted during the attack in Vienna on the 2nd of November 2020. The local police posted continuous reminders that people should not share video or images of the act, but instead send them straight to the authorities through a webpage administered by the police.[8] Whether this had any effect is currently unknown. However, during the Brussel lockdown in 2015, users on twitter flooded a hashtag with cat images in support of ongoing police operations in the city, an act later acknowledged as helpful by the authorities.[9]

An interview study with 44 convicted Islamists showed that content seen on the internet can be one contributing factor in radicalising individuals, and as such lead the individual to conduct a terror act by herself (either alone or with a group). This occurs especially when online propaganda correlated with news reporting, further consolidating the individual's perception of the society.[10] With continuous access to online news agencies there exists an unlimited number of (legitimate and ill-legitimate) sources to correlate with an extremist narrative. One should also not forget the opportunity of extremist groups on each end of the ideological spectrum to feed off each-others "alternative news", and as such confirming each other's prejudice views of the world.

The power of influence that internet and social media have on the radicalisation of "lone wolves" is being widely discussed through academia and society. A common perception is that an individual can become a radical by consuming radical content online, without being in contact with anyone else in the process. It has however been contested if radicalisation is even possible without the physical contact between individuals,[11] and that it is not online content as such that influence individuals to become radicals. Rather it is the behaviour (as actively searching for violent content like beheading videos) and a moral disengagement from society that has this influence.[12] Nevertheless, there is an argument to be made that the internet, and platforms like YouTube, enable radicalisation with their recommendation-algorithms which in turn enable the user to explore content that she might not have found otherwise.[13]

There is no shortage of examples where "lone wolves" used the internet as a medium to spread their world views before conducting an attack.[14] Recent events being Utøya, Norway in 2011, a 17-year-old attacking a Swedish school in 2015 and Christchurch in 2019 where all three attackers seem to have been radicalised through the computer. In the latter case the attacker claimed to have been inspired by the two former.[15] Nevertheless, the Von Behr et al argument regarding the need for at least some interaction online for the radicalisation process to progress cannot be disproven by the literature.[16] In the previous cases mentioned there had been some interaction between the attackers and other individuals online, even though these individuals did not actively (or knowingly) participate in the radicalisation.

Detecting extremist content through linguistic analysis

All communication between individuals is based on a common language. Whether spoken or written the sender and receiver need to share a common understanding for the language used. Hence, for one to be able to detect radicalisation online one must be able to detect the language used by extremists. The introduction of web 2.0 emphasised the user's role in creating content and collaborating with other users in sharing content on social networking sites. As such, users moved from internet forums towards platforms like YouTube, Facebook, and Twitter. Ergo, the user was no longer restricted by moderators in expressing her thoughts on a particular internet board. With the new platforms, moderation was instead left to the users, reporting content that they found inappropriate, and the opportunity was given to the users to write more "freely". Even though the users changed platforms for expressing their thoughts and opinions, the main method of delivery was still through text in various formats, such as "tweets" and "status updates". Thus, the use of linguistic detection of radical content to detect extremists on various web forums can still be utilised even on social media platforms.

By using linguistic pattern detection there is a possibility to detect individuals in the process of becoming radicalised. Additional challenges are however covert propaganda that on the surface seems legitimate, addressing everyday questions but aiming to entice adolescents,[17] or users who adapt to a "neutral writing style" where interpretation is left to the reader.[18] This, together with language being dynamic, as in words can change meaning depending on the sender and receiver, emphasises the need for an equally dynamic method of detection.

Detection of extremist content is possible by analysing radical users and the common phrases they use,[19] which requires individuals that have knowledge about the specific language used.[20] Consequently, there is a need not only for systems that are adaptable to an ever-changing environment, but also an understanding that not all individuals sharing radical content with a malicious intent will be subjected to removal from platforms due to "benefit of a doubt". This supports the argument for keeping a human in the loop and not leaving the decisions entirely to algorithms--the latter which have been, and continue to be, exploited by malicious users to silence critics.

Artificial intelligence and machine learning as tools to detect extremist content

The amount of content on the internet grows for every hour, and as such it is impossible for humans to detect, classify and act on radical content within a reasonable timeframe. So, there is a need to borrow the computing power from machines. As the algorithms on various platforms gets more advanced, so does the need for computers that are able to adapt to the new circumstances. Also, how one user acts in cyberspace may need to be cross-referenced and analysed together with other sources of information. Fusing open-source information from social media platforms with criminal records may help law enforcement agencies detect potential "lone wolf" terrorists.[21] Hence, artificial intelligence and machine learning have an important role to play today and in the future. Nonetheless, differentiating between legitimate and radical content may, as such, require a wide range of tools, both computer and human-based, to not falsely flag, block or remove legal content. Thus, there is an inherent risk that automatic removal without a human reviewer in the loop may lead to the removal of content from human rights groups trying to document crimes or hinder law enforcement and national intelligence and security agencies from gathering valuable intelligence.[22]

There is also a discussion to be had regarding integrity and overall user privacy. The collection of user generated content may involve users that are not suspected of (committing) any crimes. Thus, it may be illegal in some countries to even collect and store such data for any length of time by law enforcement. Although there is an opening for a common law within the European Union requiring operators to delete reported extremist content within one hour while at the same time excluding content published for research or educational purposes. This in turn means that there is a need for continuous ethics and judicial discussions on where to draw a clear line that does not risk researchers or ordinary citizens getting caught in the crossfire.

Restricting access to radicalising content

During the early 2000s, the internet was seen as a "lawless" realm where the users were free to reign and form their own world. The internet did not know national boundaries and regulations were loose (if at all existing). The infrastructure as such has been regarded as a vital part of free speech and attempts to force internet providers in moderating content has been met with resistance. Instead, the regulation has been, with few exceptions, conducted by the users themselves on forums, in chatrooms and so forth. This in turn led to various ideological corners being formed where like-minded people could discuss topics without having other people disturbing the peace. During this period, the era of filesharing began where anonymous users could share music and movie-clips through special software.

With this development the internet started to experience the sharing of audio and film clips through "hubs" of users. Before, a user needed to get invited and provided with a password to be able to connect to the other users. However, the introduction of social media platforms connected users and communities more effectively as the platforms' algorithms developed, turning the platforms into amplifiers for extremist content,[23] such as imagery or videos, where religious and political extremist groups have fully embraced the opportunities given to them, and encouraging sympathisers to conduct violent acts.

In order to restrict the access to radical content, platforms like Twitter, YouTube and Facebook have become targets where governments demand that the corporations systematically remove radical content from their servers. Driven by this, the corporations have created a shared database for hash-matching[24] that helps detect and remove (or at least unlist[25]) unwanted content.[26] The backside of this development was noticed when practitioners documenting human rights violations began to observe a shift from violent content being online "for ages" to instead being gone within one hour of publication.[27] As such, the term "internet never forgets" seems to have become a false statement that depends on whether someone manages to download content or not before the moderation system removes it.

Reed et al put forward four general recommendations for not amplifying extremist content on social media platforms:

1) removing problematic content from recommendations,
2) ensuring recommendations are from quality sources and providing users with more context and alternative perspectives,
3) Greater transparency, and
4) further research.[28]

However, terms like "problematic content" and "quality sources" may become too subjective and change over time, especially when "quality sources" like established media outlets and their connection to political power are an established ideological narrative.[29] Providing users with "alternative perspectives" may result in promoting perspectives that may not be radical per se but offer a perspective that is harmful in the long run.[30] As such, caution is needed when moving forward with restrictions on social media platforms. On one hand there is a risk of being too restrictive and as such becoming an issue of free speech on the internet, while on the other hand not being proactive may continue to amplify radicalising content.

The crack-down on extremist content has in turn led to extremist groups moving on to encrypted platforms like Telegram where the insight from the outside world is limited.[31] Conway et al concludes that the movement from Twitter to Telegram indeed sliced the numbers of users exposed to extremist content but may also create an "echo chamber"[32] since the control over user access is now greater than before.[33] The large number of "bots and channels" being blocked by Telegram[34] indicates that extremists' users have found a platform they prefer. As such, extremist groups will find alternative channels to broadcast their content when one platform starts to restrict their content, as well as spreading their content on several platforms in order to maintain a powerful image.[35]

This means that the content gets harder to find, restrict and trace due to encryption and fragmentation. However, this also implies that it becomes more difficult for followers of these groups to find radicalising content. Just like the filesharing community mentioned in the beginning, the "community" of extremism continuously adapts to the changing environment with the access to new technology. The introduction of blockchain, cryptocurrency and peer-2-peer-encryption opens for decentralised social networks where content cannot be removed by the provider and tracing users becomes harder.[36] Consequently, the ongoing fight to remove extremist content from the internet may have to shift focus from restricting content into developing offensive tools which enable the ability to shut down social networking nodes.

Neumann discusses different strategies to limit the spread of radicalising content on the internet from a U.S. perspective.[37] The author states that it is impossible to remove all extremist content from internet with terabytes of data being transferred every minute, and that even complex systems like "The great firewall of China" struggle with removing content in near real time.[38] The ongoing protest in Belarus shows that not even restricting the overall internet access within a country can fully hinder sharing of images and video. Such a strategy would also be noticed by other countries and cause concern by, at least, the population. Although removing IS-accounts on Twitter has resulted in disrupting the group's activities on that particular platform, there is a lack of empirical data available on whether such disruption operations have any long-lasting effect in limiting the exposure of radicalising content on a strategic level.[39]

All in all, there is a conflict between protecting free-speech laws (that differs between nation states) and the desire of hindering individuals from posting radicalising content online. Ethically there is a difference between detecting content and the users that disseminate radical content and restricting their access to various platforms or even the internet as such without involving the justice system. Also, data presented by Von Behr et al shows that only a small amount of reported content was actually removed by law enforcement in the United Kingdom in 2012.[40] This implies a disconnect between real world policing and the policing conducted online.

Countering the narrative

It lies in the extremist groups' interest not just to be able to share their narrative unhindered, but also be able to control and shape it to their benefit. Yet, it is not sufficient to just send out messages, or even start an argument with extremist supporters, and claim to be countering a narrative. A narrative should be understood as something more complex than just the message by itself. Ruston defines the term "narrative" as "systems of stories structured in such a way as to make meaning about the world around us". It could also be understood as a cognitive process in which the receiver structures the information in cause, effect, and consequence.[41] Hence, the challenge in countering a narrative is not only in addressing a statement or limiting the spread of content on social media. There within lies a greater challenge in influencing the cognitive process (cause, effect, consequence), which in turn requires various methods and techniques.

Braddock suggests that even though violent extremism as such is unusual, the methods used to persuade an audience may not be any different from how politicians work.[42] Neumann argues that one important measure is overall to reduce the demand for radical and extremist messages, which may be accomplished by education, mainly of young people, confronting and discrediting narratives.[43] Both studies imply that society plays a critical role in not just educating young people in the classroom, but also inoculating them to become resilient towards radicalisation. There may also be a need to analyse how media communicate around certain events, and which effects they may have on the overall narrative.

According to Baugut and Neumann radicalisation can be halted (and even reversed) when news media that clearly differentiated between Islam and terrorism was exposed to the radicalised individual.[44] As such, there may be a need for gathering information about the radicalising narrative and comparing that to the mainstream media's overall message since it may be correlating and with that confirming the individual's perception of the world. However, as the authors also point out, this requires that the individual be exposed to these types of messages. A common ground for both religious and political extremism is to quickly tie the individual to the group and shield that person from the outside world, which decreases the chance of a person being reached by a "de-radicalising" message.

Lewis claims that online personalities using YouTube as a platform and inviting radical individuals to their channels, in fact helps nurture and spread radicalising narratives.[45] This means that the popular notion in that every perspective should be given equal opportunity to speak their opinion may in fact be counter-productive and instead lend credibility to the radicalising narrative, at least in a social media context. Nevertheless, this should not be interpreted as a proposition that journalists should start lying, or not reporting events that may feed into a radical narrative. Such a suggestion would not only be unethical but also undermine the credibility of independent media and as such the foundation of democracy. Thus, self-censoring would be counter-productive and further play into the hands of the extremist narrative.

The news media does play a role when it comes to antagonising radicals, ridiculing their ability to conduct acts of terror.[46] The act of using humour to belittle extremists is not uncommon and may take many forms on the internet through memes, taunting videos, or songs. It may be performed by "informal" actors like private citizens lashing out against radical narratives that they do not approve.[47] This type of act may be based on the individuals own strong ideological position, and as such may only reach users that already share that view.[48] However, informal actors may prove to be an invaluable asset in countering narratives due to their own strong beliefs, a form of crowd-sourced detection and counter-narrative tool. These individuals may even be able to reach into forums that require an account or encrypted channels that are potential gateways to extreme content, and through that access influence individuals at risk of becoming immersed in radical content, thereby offering an alternative group for the individual to feel a part of, or at least a different perspective on current topics.[49]

There exist several examples where governmental programs have failed in being able to show efficacy in turning people away from extremism and highlighted that these types of programs may have marginalised Muslim populations. Strategies in countering narratives (and as such radicalisation) need to be grounded in research where data can be brought forward to support claims, which seems obvious in writing but somehow seems harder to implement in practice. Since the image (both in the meaning of appearance and as a graphic representation of an object) seems to be of great importance for extremist groups, an overall weakness with the reviewed literature is that imagery (and the narrative it carries) is not discussed at any great length. It stands to reason that when communication occurs through carefully produced videos or elaborately designed photos, the countering of that communication needs to adapt and transform into the same design, at least visually.

Conclusion

The effectiveness of counter-radicalisation schemes has been questioned throughout the years. The literature reviewed in this article is just a snapshot of an ever-growing research field that also is diverse in nature. As such, not only academic researchers are interested in the field, doing studies, and writing papers. Even though the majority of articles used in this review derives from journals, there lies a value in also acknowledging "white papers" from researchers and practitioners connected to various institutes as well as tech companies. Firstly, due to the fast-paced nature of internet and technology may render studies obsolete within years or even months. As such "white papers" may provide insights, explaining current trends and point towards possible future developments within a specific field of study. It should be noted that these papers, even though not peer-reviewed, still rest on a scientific methodology when referencing to specific events as well as being reviewed by subject matter experts before publication. Secondly, "white papers" provide a perspective on how future policies adopted by governments could look on a practical level, thus, explaining why academia also from time to time refers to these papers when gathering information and empirical data.

The area of social media as a radicalising instrument is of great interest for not only governments but also the companies that provide the platforms and the society that share the cyberspace with extremists. Radicalisation is a process involving not just the individual and the content, but also the method of delivering the content (internet and social media) and a narrative (the information structure of cause, effect, and consequence). Hence, there is a need for a holistic approach in preventing radicalisation online, not just focusing on restricting the content or blocking users. With every restriction the extremists find either a new way of disseminating their content to followers on the current platform or move on to a new one, circumventing the filter.

This means that the feeling of victory for the society will be short lived and adds to the dangerous illusion of handling the issue of online radicalisation, especially in the future when extremist groups may move on to de-centralised platforms enabled by blockchain and peer-2-peer technology. This will require state and non-state actors to develop tools that can assist in detecting and disabling these platforms. The need for quick and accurate detection opens the doors for machine learning and artificial intelligence becoming vital in process of detecting and countering radicalising content. However, a human in-the-loop is still required, not only to add context knowledge due to changing language but also detecting and validating the use of "neutral" language used by certain users and as such offering a "second opinion" to the machine.

Overall, the sharing of both violent and non-violent extremist content relies on there being a demand. When an individual morally disconnects from society a threshold is passed in the process of becoming radicalised. As such, informal actors, such as private citizens and independent organisations, can help in countering radicalising narratives as well as picking up individuals that are on the path to becoming extremists. Nevertheless, societies and the educational systems play an important role in educating and inoculating young people and sensing when an adolescent may be at risk. This in turn may help in building cognitive resilience when the individual encounters a radical narrative online. However, caution is warranted when the narrative focuses on the state and media acting maliciously in conjunction with each other. Moderation may also become more difficult as spreaders of extremist content may adopt a "neutral" style of language which could be used against social media platforms under the pretext of free speech. This also means that removing content or restricting access to platforms further adds to the narrative, confirming what the individual thinks he or she already knows.

There is no way of knowing how social media will develop, or where the majority of internet users will spend their time in the future. It is however highly likely that extremist groups and radicalising content will be present wherever maximum exposure can be guaranteed. Thus, there will be a continuing need to study how users interact (talk, share information etc) on the internet and social media, and how extremist groups use new technology within the realm of cyberspace.

Suggestion for future research

Though studies suggest approaches like moderation or using informal actors when countering radicalisation online, there is a gap within the field in measuring if, and to which extent, these recommendations have any effect. This gap derives from a lack of available empirical data on whether disruption operations have any long-lasting effect in limiting the exposure of radicalising content on a strategic level on social media platforms. A gap that may be hard to close due to complications regarding how to measure "success" within this subject area since such success may be operationalised differently by different state actors. However, with indications that the process of radicalisation can be reversed there is a need to probe that area thoroughly.

Since informal actors, such as private citizens, can directly interact with radical users online there is a need for examining if these interactions raise the risk of further spreading radicalising content. Also, there is a need to investigate if these interactions, especially those which ridicule extremists, carry some risk of physical harm to the individuals interacting with radical actors. It is possible to draw conclusions from citizens' engagement with organised crime such as Mexican cartels which have shown themselves more than able to locate and harm twitter users within the country. Ergo, if citizens risk being harmed due to social media activity there is a need to identify such risks and adopt tools to prevent violent acts.

With new technology being presented in a rapid pace, there is a need for further cooperation between academia, government agencies and the practitioners (i.e. tech companies) in identifying potential areas of interest for extremists. Particular areas of interest are the use of new platforms for communicating, technology that supports the dissemination of material or even the development of language. Understanding the interests of extremists in general could act as a method in detecting individuals or groups at risk of becoming radicalised.

The literature reviewed has a clear focus on locating radicalising content on the internet, and detecting users spreading that content on various platforms. Consequently, it seems clear that there exists a solid ground both methodically and technologically how to proceed in finding content and users. Future research needs to be concentrated elsewhere such as understanding how narratives are being crafted online and how a single post going viral on social media can influence an individual to commit violence.

All in all, research findings within this area need to be brought forward to the public and politicians, enabling a debate about internet regulation, privacy, and free speech as well as policymaking based on research. The future of (online) democracy may depend on it.

The author serves as a Sergeant first class in the Swedish armed forces and is currently completing a master's degree in political science.

Notes

[1] Baugut, Philip and Neumann, Katharina: "Online propaganda use during Islamist radicalization", *Information, Communication & Society,* 23:11, 2020, pp. 1570-1592.

[2] Braddock, Kurt: "Vaccinating against hate: Using attitudinal inoculation to confer resistance to persuasion by extremist propaganda", *Terrorism and Political Violence*, 2019, pp. 1-23.

[3] Gaudette, Tiana; Scrivens, Ryan and Venkatesh, Vivek: "The Role of the Internet in Facilitating Violent Extremism: Insights from Former Right-Wing Extremists", *Terrorism and Political Violence*, 2020, pp. 1-18.

[4] Von Behr, Ines; Reding, Anais; Edwards, Charlie and Gribbon, Luke: *Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism,* Rand, Brussels 2013.

[5] Lewis, Rebecca: "Alternative influence: Broadcasting the reactionary right on YouTube", *Data & Society,* 18, 2018.

[6] Mair, David: "# Westgate: A case study: How al-Shabaab used Twitter during an ongoing attack", *Studies in conflict & terrorism,* 40, No. 1, 2017, pp. 24-43.

[7] Flynn, Meagan: "No One Who Watched New Zealand Shooter's Video Live Reported It to Facebook, Company Says", *The Washington Post,* 2019-03-19, *https://www.washingtonpost.com/nation/2019/03/19/new-zealand-mosque-shooters-facebook-live-stream-was-viewed-thousands-times-before-being-removed/*, (2021-01-26).

[8] Polizei Wien, *https://twitter.com/LPDWien/status/1323364760260411392*, (2021-01-26).

[9] "Belgians Tweet Cat Pictures during #Brussels Lockdown", *BBC News*, 2015-11-23, *https://www.bbc.com/news/world-europe-34897645*, (2021-01-26).

[10] Op. cit., Baugut, Philip and Katharina Neumann see note 1, p. 1446

[11] Op. cit., Von Behr, Ines; Reding, Anais; Edwards, Charlie and Gribbon, Luke, see note 4.

[12] Frissen, Thomas: "Internet, the great radicalizer? Exploring relationships between seeking for online extremist materials and cognitive radicalization in young adults", *Computers in Human Behavior,* 114, 2020.

[13] Op. cit., Lewis, Rebecca, see note 5.

[14] Cohen, Katie; Johansson, Fredrik; Kaati, Lisa and Clausen Mork, Jonas: "Detecting linguistic markers for radical violence in social media", *Terrorism and Political Violence*, 26, No. 1, 2014, pp. 246-256.

[15] According to the manifesto attributed to the Christchurch attacker.

[16] Op. cit., Von Behr, Ines; Reding, Anais; Edwards, Charlie and Gribbon, Luke, see note 4.

[17] Op. cit., Lewis, Rebecca, see note 8.

[18] Åkerlund, Mathilda: "The importance of influential users in (re) producing Swedish far-right discourse on Twitter", *European Journal of Communication*, 35, No. 6, 2020, pp. 613-628.

[19] Rekik, Amal; Jamoussi, Salma and Ben Hamadou, Abdelmajid: "A recursive methodology for radical communities' detection on social networks", *Procedia Computer Science*, 176, 2020, pp. 2010-2019.

[20] Conway, Maura; Khawaja, Moign; Lakhani, Suraj; Reffin, Jeremy; Robertson, Andrew and Weir, David: "Disrupting Daesh: Measuring takedown of online terrorist material and its impacts", *Studies in Conflict & Terrorism*, 42, No. 1-2, 2019, pp. 141-160.

[21] Hung, Benjamin WK.; Jayasumana, Anura P. and Bandara, Vidarshana W.: "INSiGHT: A system to detect violent extremist radicalization trajectories in dynamic graphs", *Data & Knowledge Engineering*, 118, 2018, pp. 52-70.

[22] Banchik, Anna Veronica: "Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights-related content", *New Media & Society*, 2020.

[23] Reed, Alastair; Whittaker, Joe; Votta, Fabio and Looney, Sean: "Radical Filter Bubbles: Social Media Personalization Algorithms and Extremist Content", *Global Research Network on Terrorism and Technology,* London 2019.

[24] Every uploaded object generates a hash. No matter if a user tries to re-upload, the hash will be unchanged if the file itself is not modified. That hash-string can be matched against other strings in a database.

[25] Unlisting content means that it will not be searchable or show up in any recommendations. The unlisted content remains on the platform for the uploader to see. This enables the uploader to correct the content according to pol-

icy, and the platform to keep content accessible for employees during further investigation.

[26] Op. cit., Conway, Maura; Khawaja, Moign; Lakhani, Suraj; Reffin, Jeremy; Robertson, Andrew and Weir, David, see note 20.

[27] Op. cit., Banchik, Anna Veronica, see note 22, p. 6

[28] Op. cit., Reed, Alastair; Whittaker, Joe; Votta, Fabio and Looney, Sean, see note 23, p. 15, 16.

[29] Op. cit., Lewis, Rebecca, see note 5, p. 4.

[30] Although outside the scope of this article, the debate regarding vaccination is one example where presenting "alternative perspectives" may have caused harm to the society.

[31] Bloom, Mia; Tiflati, Hicham and Horgan, John: "Navigating ISIS's preferred platform: Telegram1", *Terrorism and Political Violence*, 31, No. 6, 2019, pp. 1242-1254; Op. cit., Conway, Maura; Khawaja, Moign; Lakhani, Suraj; Reffin, Jeremy; Robertson, Andrew and Weir, David, see note 3, p. 13, 14.

[32] An environment in which somebody encounters only opinions and beliefs similar to their own and does not have to consider alternatives. Retrieved from *https://www.oxfordlearnersdictionaries.com/us/definition/english/echo-chamber.*

[33] Op. cit., Conway, Maura; Khawaja, Moign; Lakhani, Suraj; Reffin, Jeremy; Robertson, Andrew and Weir, David, see note 20.

[34] Total number of terrorist bots and channels blocked in 2020: November: 17975, October: 18431, September: 16859. Data fetched 10 December 2020 from the official telegram channel, ISIS Watch.

[35] Yarchi, Moran: "ISIS's media strategy as image warfare: Strategic messaging over time and across platforms", *Communication and the Public*, 4, No. 1, 2019, pp. 53-67.

[36] Mott, Gareth: "A storm on the horizon? "Twister" and the implications of the blockchain and peer-to-peer social networks for online violent extremism", *Studies in Conflict & Terrorism*, 42, No. 1-2, 2019, pp. 206-227.

[37] Neumann, Peter R.: "Options and strategies for countering online radicalization in the United States", *Studies in Conflict & Terrorism*, 36, No. 6, 2013, pp. 431-459.

[38] Ibid, p. 439.

[39] Op. cit., Conway, Maura; Khawaja, Moign; Lakhani, Suraj; Reffin, Jeremy; Robertson, Andrew and Weir, David, see note 20.

[40] Op. cit., Von Behr, Ines; Reding, Anais; Edwards, Charlie and Gribbon, Luke, see note 4, p. 5.

[41] Ruston, Scott: "Narrative in strategic communications" in Lange-Ionatamishvili, E. (ed.): *Russia's footprint in the Nordic-Baltic information environment*, Ryga: NATO Strategic Communications Centre of Excellence, 2018, p. 16.

[42] Op. cit., Braddock, Kurt, see note 2, p. 2, 3.

[43] Op. cit., Neumann, Peter R., see note 37, p. 433.

[44] Op. cit., Baugut, Philip and Neumann, Katharina, see note 1, p. 1453.

[45] Op. cit., Lewis, Rebecca, see note 5, p. 36.

[46] Op. cit., Baugut, Philip and Neumann, Katharina, see note 1, p. 1452.

[47] Lee, Benjamin: "Countering violent extremism online: The experiences of informal counter messaging actors", *Policy & Internet*, 12, No. 1, 2020, pp. 66-87.

[48] Ibid., p. 83.

[49] Op. cit., Gaudette, Tiana; Scrivens, Ryan and Venkatesh, Vivek, see note 3, p. 8, 9.