Combating radicalization by algorithm

Many countries have formulated strategies to combat radicalization, and regulating and monitoring recommendations on social network platforms is part of these strategies. This article explores the concept of radicalization and the proposed causal link between recommendations and radicalization, highlighting some of the key characteristics of recommendation engines. Based on this background potential risk mitigation measures can be identified that directly affect the recommendation engines. To conclude, some key problems and dilemmas for risk management are discussed.

Dr. Alexander Boer
is a senior manager at
KPMG Responsible AI.

INTRODUCTION

Recommendation systems, a subclass of information filtering systems that helps us choose information items from an overwhelmingly large collection of them, are everywhere. We all use them. All the time. Every time we use a large Internet platform for shopping, for preparing travel, for keeping in touch with our social network, for filtering the news we read, or for finding the answer to a question we have. All large Internet platforms filter, rank, and present content using a recommendation engine. Proactively, or as a reaction to our actions, like entering a search phrase or question.

Why they are important – from the point of view of the end user at least – is captured well by an infamous and prescient quote of half a century ago:

> "What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."
>
> – Herbert Simon, 1971

The recommendation engine is a technology that, in a world of too much information, helps us to structure our information environment so that we can spend our limited attention on the information we want. Clearly, recommendation engines matter. And increasingly, we criticize how they work. Or how they don't work well enough. Or safely enough. Recommendations based on profiling may for instance reveal the user's sensitive characteristics, like sexual orientation. Online platforms may abuse recommendation engines for self-preferencing: boosting the visibility of information about their own products or services or those of close affiliates, over those of others. And they can be manipulated. Malicious users constantly invent new ways to bias or sabotage recommendations. And, finally, recommendation engines are often claimed to play a major role in the spread of false information ([Whit21]), the radicalization of people caught in harmful filter bubbles, and in that way may contribute to terrorism and other harms to democratic society. And they can harm us in these ways because they effectively control our attention, and we can't really do without them.

In response, governments increasingly take action to regulate recommendation engines, and impose requirements relating to privacy, fairness, safety, transparency, and resilience of recommendation mechanisms. But operationalizing these requirements into practical solutions for managing risk & control, or for balancing these risk and control options against the legitimate business interests of the platform, is a major challenge. Just like auditing such solutions, and neutrally and objectively reporting findings about them.

To investigate the nature of this risk management challenge, we will focus on the most ominous among the example risks listed earlier: What can an online platform realistically do to reduce the risk that – specifically – its recommendation engines contribute to the radicalization of people? Furthermore, we will specifically adopt a risk management and audit perspective on the problem: How does one judge the impact of the recommendation engine on radicalization? And can one reasonably mitigate the risk through interventions in the recommendation engine? Or are other solutions more effective?

To investigate this problem, we will first dive into the proposed role of social media platforms and filter bubbles in radicalization, and into the notion of radicalization, radicalized persons, and radicalization processes. Then we move to the use of reinforcement learning in recommendation engines and its weaknesses in relation to those filter bubbles. And, finally, we look at potential risk mitigation measures that directly address recommendations, arriving at some key problems and dilemmas for risk management.

ONLINE RADICALIZATION

Policy makers have expressed repeated concern over the role of social media in the radicalization process in the last two decades. Recent strong suspicions of weaponization of radicalization by Russia through large-scale use of bots to spread and then upvote radicalizing content in order to manipulate recommendation systems ([Geis23]), made the issue only more pressing. Adding to the sense of urgency are the ever more convincing deepfakes created by Generative AI, making it harder and harder to take down bot farms.

Many countries have formulated strategies to combat radicalization, and regulating and monitoring social network platforms is usually part of these strategies. One example is the requirement of a recent European regulation, the Digital Services Act, which entered into force in the spring of 2023, to assess and mitigate systemic risks caused by a platform's recommendation engines. Contributing to radicalization is obviously one of those risks to be assessed and mitigated. The implementation of risk management is subject to a yearly audit.

It is a very contentious issue as well. Recently, for instance, a number of organizations filed a complaint before the French administrative court, the Conseil d'État, against the French decree implementing another European regulation that combats the dissemination of terrorist content online (TERREG; 2021/784). The organizations are requesting a ruling from the Court of Justice of the EU on the validity of this regulation in light of human rights.

Given that the Internet plays an important role in daily life, it is unsurprising that radicals exist on the web ([Bens13], [Neum13]). Studies show that the Internet is important to radicalization ([Behr13], [Bast18]), and that its importance is growing ([Gill17], [Jens18]), but not necessarily at the expense of offline factors ([Behr13], [Hera21]). Both jihadist and far-right groups have deployed strategies to spread content for a long time. The Islamic State, for instance, developed content persuading young individuals to engage in terrorism or travel to Iraq or Syria. The group at one point had a substantial reach on Twitter, and a presence on up to 330 different platforms ([Berg15]). The far right has a very substantial online presence on major social media platforms, and even more on fringe platforms, which are less likely to remove extreme content ([Conw19]). In both cases, governments have sometimes tried to disrupt activities, but communities apparently have no problem finding platforms to spread content ([Fish19]).

Although there is a substantial amount of research into the reach of extremist propaganda on the Internet, there is still little actionable insight into how and when this radicalizes people ([Ruddoo]). There is some evidence that engaging with propaganda generally increases support for radical groups ([Reev19]), and that priming on existential threat or uncertainty increases its persuasiveness ([Rieg13]). But radical extremist content clearly does not radicalize an indiscriminate part of the audience reached ([Sage14], [Arch15], [Aly17]). Which brings us to the question of how the radicalization process works.

RADICALS AND THE RADICALIZATION PROCESS

In common usage, the word *radicalization* is ambiguous: it is either developing increasingly extreme beliefs, engaging in increasingly extreme actions (such as terrorism), or becoming part of a community of radicals. Moreover, there has been little success at attempting to profile or model a radicalized individual. The research populations would be difficult to reach ([Jens18]). The literature which attempts to do so, is unable to account for observed communities ([Boru17]), and evidence relies on vague selection criteria ([Jens18]).

Empirical research tends to focus on risk factors for radicalization, either internal vulnerabilities of the person or external factors such as life stressors. These factors are not necessary or sufficient for radicalization, but appear at higher rates in radicalized populations than in the general population.

Many countries have formulated strategies to combat radicalization

Some external factors stand out. So-called *radicalization hotspots* produce more radicalized individuals than would be expected ([Varv16], [Vidi17]). Some environments encourage or fail to suppress radicalization. Engagement with criminal activity has a clear relationship with radicalization – often referred to as the *crime/terror nexus*: the personal needs and desires of criminals appear to be similar to those of terrorists ([Basr16]).

Research on internal factors presents a mixed picture:

- Demographically, radicalized individuals differ somewhat from the general population, but a lot less than one might expect ([Vidi17]). But people that engage in extreme action as a result, are clearly male (ranging from 85-95%) and young (mid-twenties) ([Horg16]).
- Socioeconomic background is often mentioned as a potential stressor, but research on this topic is mixed ([Schm13]). Some studies find this to be a poor explanation ([Reyn17]), while others find a small correlation ([Cruz18]).
- Education is not a factor: some radicalized populations are well-educated, and others poorly educated ([Gill15]).
- On mental health, results are mixed as well: individuals that act as part of a group do not stand out ([Horg08]), but lone actors do have a higher than average prevalence of mental health disorders ([Corn16]).

External factors and internal vulnerabilities interact ([Clem20]). The individual chooses how to spend their time (*self-selection*), and there is the setting they were born and raised into (*social selection*). Self-selection will play an important role when we look at the role of recommendation engines in radicalization. In any case, we don't get a very clear *profile* of the vulnerable community or person, calling into question whether platforms can effectively tailor measures to vulnerable individuals.

SOCIAL MEDIA PLATFORMS AND THE FILTER BUBBLE EFFECT

One of the key concerns in the online radicalization policy debate is whether platforms are *amplifying* radical extremist content towards social media users ([Coun20]), creating a *filter bubble effect* ([Pari11]). In a meta-review on this topic, 8 of 11 studies identified suggest that search and recommendation algorithms *may* amplify extremist content ([GIFC21]). Another review of the existing literature came to a similarly careful conclusion ([Knot21]). But both studies identify basic methodological flaws and limitations in the academic literature, like not employing control groups or experimental conditions and a focus on platforms that are easiest to study as a black box.

More recently, studies have started to address these limitations ([Husz22]). But the extent to which machine learning algorithms used for search and recommendation of content play a role in the user journey towards radicalization, is hard to quantify in isolation, because the effects of recommendation cannot be separated from the "self-selection" choices the user of a social media platform makes towards personalization of content: the user searches for content, and joins online communities. Let us call this factor *self-selected personalization*, as opposed to *algorithm-selected personalization*, where the recommendation system makes you part of a virtual community. An ideal study design would reproduce the *user journeys* of users vulnerable to radicalization in a recommendation-driven environment. Of course, that solution depends on accurate profiling for empirical validation.

THE ROLE OF RECOMMENDATION ENGINES

Business models for social media platforms usually rest on user engagement with the platform. Holding and steering the attention of the user is a necessary aspect of the business model. Social media platforms provide services, free or at low cost per user, through which individuals, communities, and organizations that use those services can access, share, co-create, discuss, and curate content. Broadly speaking, the business model of the platform in some way always depends on its ability to proactively offer engaging content. The objective of the recommendation engine used by a platform is to maximize user engagement with content offered by the social media platform.

Moreover, engagement – the user's voluntary interactions with content on the platform – is the only thing that can be more or less objectively measured by the platform. One cannot fully rely on user ratings expressing preferences for content alone, due to the obvious vulnerability of ratings to straightforward manipulation by content creators.

User engagement with content creates the metadata that can be used to gain insight into individual user preferences for content, and the generic engagement qualities of the content. Some of this insight is intentionally and consciously created by users (*self-selected personalization*), and some of this insight is unintentionally created by users as they reveal their preferences by way of their actions (*algorithm-selected personalization*, [Bodó18]).

An example of self-selected personalization is joining a channel on YouTube or a sub on Reddit. The user is aware of having *expressed* a preference for content with certain characteristics by classifying himself or herself. Algorithm-selected personalization, on the other hand, takes place if metadata is used to *predict* user preferences for content, without the user ever consciously expressing a preference. In this mode, every user action is interpreted as a choice from a limited menu of options, *revealing* his or her preferences.

The user engagement data obtained through these two types of user interaction, is used by the social media platform's recommendation engines to filter, rank, and present content, or to provide search suggestions. Through control of suggesting, filtering, ranking, and presenting content to the user, the recommendation engine has a major impact on what content the user *can* easily engage with. The algorithm sets the menu of options the user expresses preference through. Using an algorithm to predict what content a user will engage with, and then using that prediction for directly *controlling* the probability of user engagement by proactively offering the content, opens up a *popularity bias* cycle ([Abdo19], see Figure 1). Known content that users are known to engage with, gets ever higher scores at the expense of content never seen. That means there will be a *cold start* problem ([Boba12]) for new content that has never been engaged with and is therefore not (yet) "popular". The new content would never be recommended. The platform must counteract the bias that would suppress new, potentially engaging, content in some way. There will be a similar cold start problem for being able to engage new users, whose preferences are unknown because they have not made enough choices yet. Furthermore, there will be a bias for assuming new users fit within the first candidate user categories identified, potentially leading to undesirable filter bubbles.

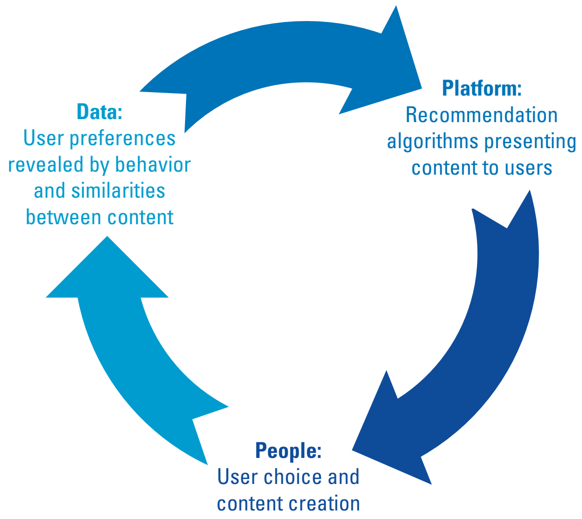
Data:
User preferences revealed by behavior and similarities between content
Platform:
Recommendation algorithms presenting content to users
People:
User choice and content creation

Figure 1. How recommendations feed popularity bias.

Both cold start problems must obviously be addressed in some practical manner by the recommendation system to be able to win new users, and to timely recommend new content. Timeliness of news is an important way for a platform to distinguish itself from competing platforms, after all. Recommendation engines therefore include specific solutions for cold start situations that play an important role in the functioning of recommendations, mainly:

- Forcing or goading new users towards self-selected personalization (e.g., through letting the user join communities or pick topics of interest) is a common method to address the cold start problem for new users. This allows treating the new user as part of a virtual community that selected the same options.
- Clustering on similarity of content characteristics to predict its engagement quality is an important method to address the cold start problem for new content. The clustering algorithms used for that purpose can usually be characterized as machine learning algorithms.

Based on similarity, the classified new content can be experimentally injected into recommendations to verify the initial hypothesis of its engagement qualities. The end user is essentially used as a guinea pig for new content, as the algorithms play with a balance between producing useful recommendations and testing new content on known users. This constant balancing act between optimization of measurable performance and experimentation with new content and new users is a fascinating artificial intelligence problem, and arguably the most interesting dimension in which platform recommendations differ in quality.

Reinforcement learning and cold starts

Recommendation engines are typically driven by algorithms that can be characterized as reinforcement learning algorithms ([Ricc11]). The algorithm is trained directly using the data that is produced in the production environment where it is fielded, and where it makes interventions, which makes this type of system more vulnerable to development of biases. Therefore, its performance on the task at hand cannot be reliably measured and reported in advance in a testing phase. Recommendation algorithms can be contrasted with supervised learning algorithms, which enter the production environment already trained with – usually – carefully evaluated and prepared training data.

In this context, a *cold start* is a situation in which an algorithm starts making interventions with real world impact immediately in an unknown environment, and only starts learning along the way as feedback data from the interventions it made starts coming in. Early "learning experiences" will have a big impact on the model the algorithm will develop, sometimes leading to increasing bias problems.

TOOLS AND METHODOLOGIES TO CONTROL RECOMMENDATION

In the recommendation infrastructure, platform organizations must consider how to develop and implement risk mitigation capabilities, amongst others to prevent radicalizing filter bubbles. The risk mitigation capabilities that can play a role, can be roughly categorized into the following approaches ([Saxe22]):

- *Classify and mark radicalizing content* to treat it differentially in recommendation, to attach banners to specific potentially harmful content, or to assign a reputation score to content ([Alfa18]). This method has been deployed widely in relation to COVID, for instance. Some research has focused on supporting this process with automated fact-checking ([Hass17]), but most of this work depends on initial discovery of new categories of questionable content by dedicated content moderation teams and careful curation of training examples. Note that the content classified as potentially harmful, *does* survive the content moderation process: it is not straightforwardly illegal or contrary to platform policy, because this would be a reason to remove it from the platform altogether. Instead, it is marked for being downplayed in the recommendations or treated differently in the platform user interface.

- *Shadow banning*, *influence minimization*, or *influence blocking* aims to identify users whose special treatment by the recommendation engine will reduce the spread of harmful content or formation of communities around it. This approach presumes that the structure of the network, such as the users and connections between users, is well-known, based on past behavior. A conceptually straightforward approach is to use *centrality measures or rankings* ([Saxe20]) to identify influential persons within a social network, but various forms of influence minimization strategies exist ([Bane20]). But it will obviously not work well on new users and newly forming communities around some topic.
- *Counter-campaigning* aims to combat the impact of content spread within a network by balancing it better with other content in recommendation. For example, in the case of fake news, several studies show that if users are exposed to both true and fake information, users tend to believe the true information and reduce sharing of fake content ([Oztu15], [Tana13]). Implementation of this solution either presumes the capability of the platform to strategically promote and demote marked content in recommendation ([Enna10]), or the capability of dedicated counter-campaigners to game the recommendation engine at least as effectively as bad faith actors do.
- *Curbing the activity of bots* may finally be effective to mitigate the risk of organized campaigns ([Geis23]), but that of course depends on the capability to accurately classify new users as bots.

There is a common theme here. All available methods do presume that either the users at risk, the users that function as radicalizing influencers, or the radicalizing content can be identified for these methods to work. The ability to accurately and timely classify new users and new content as potentially radicalizing is therefore basically the core problem.

COLD STARTS AND THE MANIPULATION OF RECOMMENDATION ENGINES

Which brings us back to the cold start: the cold start problem for new content is a central part of the radicalization dynamic. An understanding of how the recommendation engine addresses the cold start problem for content by way of similarity to known content, can be exploited by malicious users for creation and placement of engaging content targeting specific types of users ("gaming the algorithm" and "topic hijacking"). And these malicious users can create new accounts to do so.

Topic hijacking

Exploiting clustering on content similarity with low effort fakes is usually shockingly easy for breaking news topics, for instance. An example: when the war in Ukraine had just started, a Twitter message that reads as follows was briefly circulating:

Rumors have come out of Kiev that this Ukrainian soldier has killed 23 Russian soldiers defending Kiev. The Russians have dubbed her "kiss of death". Repost if you think Putin is a war criminal!

The message was accompanied by an old photo of an attractive woman wearing a military uniform of the Israeli Defense Forces (IDF). Identifying it as fake is very easy for a specific subpopulation of users: millions of Israelis and Ukrainians. But it takes some time before the message that it is fake filters through the enthusiastic endorsements. The fake works because of its placement in a very popular breaking news category, because it is "engaging" content, because it was very similar in tone to other content that became popular that day, and because neither the recommendation engine nor the average news follower is trained to tell apart military uniforms.

Accurate classification of uploaded new and original content as manipulative is hard, however motivated platforms may be in doing it. Nor will a recommendation algorithm be able to identify the user subpopulation who would immediately recognize it as bad faith manipulation, and take immediate advantage of their feedback. For that we still primarily depend on human moderators, and time.

Intentional exploitation of limitations of clustering based on content similarity can be abused to amplify radicalizing content to specific communities of users. It is hard to automatically flag it with high accuracy. This is essentially the same problem as the one of content moderation for illegal content. But is a problem that is even harder to solve, because most of the content does not meet the illegality threshold for removing content altogether or suspending the accounts of the content creators.

A DELICATE BALANCING ACT

In summary, we know from the outset that automated classification of new content as being radicalizing is a step behind current events, and is likely to suffer from low accuracy when it is most important. Automated flagging of such content will raise both many false negatives and false positives, easily overwhelming human moderation teams.

Combating radicalization through the recommendation system itself will then become a delicate balancing act:

- Low accuracy of classification of content as radicalizing and automatically acting on it, will create a general balancing problem with freedom of expression concerns, as recognized by Council of Europe guideline CM/Rec(2022)13, and discussed in detail in [Helb20].
- Profiling radicalized individuals or individuals prone to radicalization for treating them differentially in recommendation, depends on predictions that are rather hairy from a privacy and human rights perspective. The platform takes a risk, both legally and reputation-wise, by adopting a user profiling approach.
- Third-party counter-campaigning content specifically designed to reach radicalizing communities by taking advantage of content similarity may be classified as similar to radicalizing content, and for that reason risk automatic flagging for demotion. Counter-campaigners may ironically be suppressed by measures designed to take out bad faith actors.
- Classification of content as potentially radicalizing and demoting it for recommendations may be considered not fair, unreasonable, or discriminatory towards content creators, and may negatively affect the income of businesses that create the content. This would, amongst others, be at odds with requirements of the Digital Market Act, another European regulation, which requires fair and non-discriminatory treatment of content creators.

On top of these balancing problems, strategies and automated tools will have to be constantly adapted to changing characteristics of both radicalizing and highly engaging content. This requires this delicate balancing act to be performed continuously, which will take a lot of effort.

This raises a major proportionality question: How much effort do we really expect from platforms to combat radicalization through tinkering with the recommendation infrastructure? If the outcome of balancing the rights of content creators and users against the risks of radicalization raises concerns about the proportionality of intervening with recommendation, the easy way out is not to tinker with the recommendation infrastructure itself.

There are alternatives. For instance, development of adequate content moderation resources with local language capacity and knowledge of national and re-gional contexts and specificities would allow platforms to react faster to new forms of radicalizing content. This is essentially an outside-in way of working, responding to phenomena that are spotted within specific communities. And one can choose to avoid risk by preventing or limiting exposure to "risky" categories of content altogether. If you want to avoid *Plandemic*-like conspiracies (2020), you could prohibit any medical advice from unauthorized channels in the terms and conditions of the platform. Instagram's move in March 2024 to limit political news unless users explicitly opt in, potentially fits in this category.

> Expect that platforms will not focus on the recommendation mechanisms and the creation of "filter bubbles", but on other mechanisms to control content

It is reasonable to expect that platforms will not focus on the recommendation mechanisms and the creation of "filter bubbles", but rather on other, less delicate, mechanisms to control the presence of objectionable radicalizing content.

CONCLUSION

In conclusion, the realm of recommendation emerges not only as a powerful and necessary tool for guiding social interactions, but also as a potential battleground between those that attempt to manipulate algorithms to further radicalization and policy makers and platforms attempting to counteract them. Whether the algorithms *cause* online radicalization is not that clear. And neither is whether the algorithms are the right means to stop it. As policy makers and platform operators navigate the intricate web of risks and challenges posed by recommendation mechanisms, it becomes evident that a nuanced and adaptive approach is essential. By delving into the complex interplay between user engagement, content personalization, and risk mitigation strategies, stakeholders can strive towards creating a digital environment that promotes safety, diversity of content, and responsible online interactions. But the way forward is not clearcut. The journey towards effective risk management within recommendation engines is ongoing, requiring constant vigilance, innovation, and a commitment to balancing competing interests to foster a healthier online ecosystem.

References

[Abdo19] Abdollahpouri, H. (2019). Popularity Bias in Ranking and Recommendation. *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, January 2019, 529-530. https://doi.org/10.1145/3306618.3314309

[Alfa18] De Alfaro, L., Di Pierro, M., Tacchini, E., Ballarin, G., Vedova, M.L.D., & Moret, S. (2018). *Reputation systems for news on twitter: a large-scale study.*

[Aly17] Aly, A. (2017). Brothers, Believers, Brave Mujahideen: Focusing Attention on the Audience of Violent Jihadist Preachers. *Studies in Conflict & Terrorism*, *40*(1), 62-76. https://doi.org/10.1080/1057610X.2016.1157407

[Arch15] Archetti, C. (2015). Terrorism, Communication and New Media: Explaining Radicalization in the Digital Age. *Perspectives on Terrorism*, *9*(1), 49-59. http://www.jstor.org/stable/26297326

[Bane20] Banerjee, S., Jenamani, M. & Pratihar, D.K. (2020). A survey on influence maximization in a social network. *Knowledge and Information Systems*, 62, 3417-3455. https://doi.org/10.1007/s10115-020-01461-4

[Basr16] Basra, R., Neumann, P., & Brunner, C. (2016). *Criminal Pasts, Terrorist Futures: European Jihadists and the New Crime-Terror Nexus.* ICSR. https://icsr.info/wp-content/uploads/2016/10/ICSR-Report-Criminal-Pasts-Terrorist-Futures-European-Jihadists-and-the-New-Crime-Terror-Nexus.pdf

[Bast18] Bastug, M.F., Douai, A., & Akca, D. (2018). Exploring the "Demand Side" of Online Radicalization: Evidence from the Canadian Context. *Studies in Conflict & Terrorism*, *43*(7), 616-637. https://doi.org/10.1080/1057610X.2018.1494409

[Behr13] Von Behr, I., Reding, A., Edward, C., & Gribbon, L. (2013). Radicalisation in the Digital Era: The Use of the Internet in 15 Cases of Terrorism and Extremism. RAND. https://www.rand.org/pubs/research_reports/RR453.html

[Bens13] Benson, D.C. (2014). Why the Internet Is Not Increasing Terrorism. *Security Studies*, *23*(2), 293-328. https://doi.org/10.1080/09636412.2014.905353

[Berg15] Berger, J.M. & Morgan, J. (2015). *The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter*. The Brookings Project on U.S. Relations with the Islamic World: Analysis Paper, No. 20, March 2015. https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf

[Boba12] Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems*, *26*(February), 225-238. https://doi.org/10.1016/j.knosys.2011.07.021

[Bodó18] Bodó, B., Helberger, N., Eskens, S., & Möller, J. (2019). Interested in Diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. *Digital Journalism*, *7*(2), 206-229. https://doi.org/10.1080/21670811.2018.1521292

[Boru17] Borum, R. (2017). The Etiology of Radicalization. *The Handbook of the Criminology of Terrorism* (pp. 17-32). John Wiley & Sons.

[Clem20] Clemmow, C. (2020). *Risk Factors and Indicators for Engagement in Violent Extremism.* University College London.

[Conw19] Conway, M., Scrivens, R., & Macnair, L. (2019). Right-Wing Extremists' Persistent Online Presence: History and Contemporary Trends. ICCT Policy Brief, October, 2019. https://www.icct.nl/publication/right-wing-extremists-persistent-online-presence-history-and-contemporary-trends/

[Corn16] Corner, E., Gill, P., & Mason, O. (2015). Mental Health Disorders and the Terrorist: A Research Note Probing Selection Effects and Disorder Prevalence. *Studies in Conflict & Terrorism*, *39*(6), 560-568. https://doi.org/10.1080/1057610X.2015.1120099

[Coun20] Council of the European Union (2020). *The Role of Algorithmic Amplification in Promoting Violent and Extremist Content and Its Dissemination on Platforms and Social Media.* Brussels.

[Cruz18] Cruz, E., D'Alessio, S.J., & Stolzenberg, L. (2018). The Labor Market and Terrorism. *Studies in Conflict & Terrorism*, *43*(3), 224-238. https://doi.org/10.1080/1057610X.2018.1455372

[Enna10] Ennals, R., Trushkowsky, B., & Agosta, J.M. (2010). Highlighting disputed claims on the web. *Proceedings of the 19th International Conference on World Wide Web* (pp. 341-350). ACM. https://doi.org/10.1145/1772690.1772726

[Fish19] Fisher, A., Prucha, N., & Winterbotham, E. (2019). Mapping the Jihadist Information Ecosystem: Towards the Next Generation of Disruption Capability. *Global Research Network on Terrorism and Technology*, 6. Conway, Scrivens, and Macnair.

[Geis23] Geissler, D., Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). Russian propaganda on social media during the 2022 invasion of Ukraine. *EPJ Data Science*, *12*(1), 35. https://doi.org/10.1140/epjds/s13688-023-00414-5

[GIFC21] GIFCT (2021, July). *Content-Sharing Algorithms, Processes, and Positive Interventions Working Group.* Global Internet Forum to Counter Terrorism.

[Gill15] Gill, P., Corner, E., Thornton, A., & Conway, M. (2015). *What Are the Roles of the Internet in Terrorism?* Vox Pol. http://voxpol.eu/wp-content/uploads/2015/11/DCUJ3518_VOX_Lone_Actors_report_02.11.15_WEB.pdf

[Gill17] Gill, P., Corner, E., Conway, M., Thornton, A., Bloom, M., & Horgan, J. (2017). Terrorist Use of the Internet by the Numbers: Quantifying Behaviors, Patterns, and Processes. *Criminology and Public Policy*, *16*(1), 99-117. https://doi.org/10.1111/1745-9133.12249

[Hass17] Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017). Toward automated fact-checking: detecting check-worthy factual claims by claimbuster. *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1803-1812). ACM. https://doi.org/10.1145/3097983.3098131

[Helb20] Helberger, N., Borchardt, A., & Vaccari, C. (2022, 14 September). How Council of Europe guidelines on managing the impact of digital technologies on freedom of expression complement the DSA. Media@LSE Blog. https://blogs.lse.ac.uk/medialse/2022/09/14/how-council-of-europe-guidelines-on-managing-the-impact-of-digital-technologies-on-freedom-of-expression-complement-the-dsa/

[Hera21] Herath, C. & Whittaker, J. (2021). Online Radicalisation: Moving beyond a Simple Dichotomy, *Terrorism and Political Violence*, *35*(5), 1027-1048. https://doi.org/10.1080/09546553.2021.1998008

[Horg08] Horgan, J. (2008). From Profiles to Pathways and Roots to Routes: Perspectives from Psychology on Radicalization into Terrorism. *The Annals of the American Academy of Political and Social Science*, *618*(1), 80-94. https://doi.org/10.1177/0002716208317539

[Horg16] Horgan, J., Shortland, N., Abbasciano, S., & Walsh, S. (2016). Actions Speak Louder than Words: A Behavioral Analysis of 183 Individuals Convicted for Terrorist Offenses in the United States from 1995 to 2012. *Journal of Forensic Sciences*, *61*(5), 1228-1237. https://doi.org/10.1111/1556-4029.13115

[Husz22] Huszár, F., Ktena, S.I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2021). Algorithmic Amplification of Politics on Twitter. *Proceedings for the National Academy of Sciences of the United States of America*, *119*(1). https://doi.org/10.1073/pnas.2025334119

[Jens18] Jensen, M., James, P., LaFree, G., Safer-Lichtenstein, A., & Yates, E. (2018). The Use of Social Media by United States Extremists. National Consortium for the Study of Terrorism and Responses to Terrorism. https://www.start.umd.edu/publication/use-social-media-united-states-extremists

[Knot21] Knott, A. et al. (2021). *Responsible AI for Social Media Governance.* Global Partnership on Artificial Intelligence. https://gpai.ai/projects/responsible-ai/social-media-governance/responsible-ai-for-social-media-governance.pdf

[Neum13] Neumann, P.R. (2013). Options and Strategies for Countering Online Radicalization in the United States. *Studies in Conflict & Terrorism*, 36(6), 431-459. https://doi.org/10.1080/1057610X.2013.784568

[Oztu15] Ozturk, P., Li, H., & Sakamoto, Y. (2015). Combating rumor spread on social media: The effectiveness of refutation and warning. *2015 48th Hawaii International Conference on System Sciences* (HICSS) (pp. 2406-2414). IEEE.

[Pari11] Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You.* Penguin.

[Reev19] Reeve, Z. (2019). Engaging with Online Extremist Material: Experimental Evidence. *Terrorism and Political Violence*, *33*(8), 1-34. https://doi.org/10.1080/09546553.2019.1634559

[Reyn17] Reynolds, S.C. & Hafez, M.M. (2017). Social Network Analysis of German Foreign Fighters in Syria and Iraq. *Terrorism and Political Violence*, *31*(4), 661-686. https://doi.org/10.1080/09546553.2016.1272456

[Ricc11] Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to Recommender Systems Handbook. In: Ricci, F., Rokach, L., Shapira, B., & Kantor, P. (eds.), *Recommender Systems Handbook* (pp. 1-35). Springer. https://doi.org/10.1007/978-0-387-85820-3_1

[Rieg13] Rieger, D., Frischlich, L., & Bente, G. (2013). *Propaganda 2.0: Psychological Effects of Right-Wing and Islamic Extremist Internet Videos.* Wolters Kluwer.

[Rudd00] Ruddock, A. (2000). Media Effects. In *Understanding Audiences: Theory and Method* (pp. 37-71). Sage Publications. https://doi.org/10.4135/9780857020178.n3

[Sage14] Sageman, M. (2014). The Stagnation in Terrorism Research. *Terrorism and Political Violence*, *26*(4), 565-580. https://doi.org/10.1080/09546553.2014.895649

[Saxe20] Saxena, A., & Iyengar, S. (2020). Centrality measures in complex networks: a survey. https://arxiv.org/pdf/2011.07190v1

[Saxe22] Saxena, A., Saxena, P., & Reddy, H. (2022). Fake News Propagation and Mitigation Techniques: A Survey. In: Biswas, A., Patgiri, R., & Biswas, B. (eds.), *Principles of Social Networking. Smart Innovation, Systems and Technologies*, vol. 246 (pp. 355-386). Springer, Singapore. https://doi.org/10.1007/978-981-16-3398-0_16

[Schm13] Schmid, A.P. (2013). *Radicalisation, De-Radicalisation, Counter-Radicalisation: A Conceptual Discussion and Literature Review.* ICCT. https://www.icct.nl/sites/default/files/import/publication/ICCT-Schmid-Radicalisation-De-Radicalisation-Counter-Radicalisation-March-2013_2.pdf

[Tana13] Tanaka, Y., Sakamoto, Y., & Matsuka, T. (2013). Toward a social-technological system that inactivates false rumors through the critical thinking of crowds. *2013 46th Hawaii International Conference on System Sciences* (HICSS) (pp. 649-658). IEEE.

[Varv16] Varvelli, A. (ed.) (2016). *Jihadist Hotbeds: Understanding Local Radicalization Processes.* Italian Institute for International Political Studies.

[Vidi17] Vidino, L., Marone, F., & Entenmann, E. (2017). *Fear Thy Neighbor: Radicalization and Jihadist Attacks in the West.* Ledizioni. https://doi.org10.14672/67056194

[Whit21] Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1565

About the author

Dr. Alexander Boer is a senior manager at KPMG Responsible AI. He is an expert in artificial intelligence, law, and risk management, and holds a PhD degree in Artificial Intelligence & Law from the University of Amsterdam. At that university he worked as an artificial intelligence researcher for two decades, applying artificial intelligence technologies to practical and theoretical problems. Over the last few years Alexander built up a lot of experience with the implementation of the Digital Services Act and Digital Markets Act in online platform organizations.