Pericles

Policy recommendation and improved communication tools for law enforcement and security agencies preventing violent radicalization

Ethical case studies

John Guelke

Result Report

Coordinator:

Dr. Dominic Kudlacek
Criminological Research Institute of Lower Saxony
Lützerodestraße 9, 30161 Hannover, Germany
Mail: Dominic.Kudlacek@kfn.de

This project has received funding from the *European Union's Horizon 2020 research and innovation programme under* grant agreement No 740773

Document Evolution:

Version	**Date**	**Note of Modification**
V1.1	06.03.2019	First version of the report

TABLE OF CONTENTS

1 INTRODUCTION

This deliverable builds on the ethics work of Pericles Deliverable D6.2 'Ethical Considerations in Counter-Radicalisation' by discussing the ethics of counter-radicalisation in relation to the so called 'Fake News' phenomenon. Increasingly discussions of so called 'self-radicalisation' make reference to the easy availability of propaganda and misinformation online. As pointed out in Pericles Deliverable D1.2, currently only 1.9% of radicalisation prevention measures specifically target the internet. The same work also found that only 7.4% of interventions concerned resilience against online radicalisation (Kudlacek et al, 2017, 34).

The deliverable begins in part 1 by reprising the ethical framework for counter-radicalisation presented in deliverable D6.2. Counter-radicalisation is justified on the basis of similar considerations to those that justify counter terrorism – primarily the threat of serious violence. However, because the relation between beliefs, ideology, social networks and acts of violence are unclear, counter-radicalisation often takes an inclusive view of 'extreme' beliefs as well as extreme acts. The wider the focus, and more tenuous the relation to acts of violence, the more likely counter-radicalisation strategy is to be criticised as illiberal.

In section 2 'Fake News' is disambiguated for its relevance to both instances of political violence, and extreme beliefs. Section 3 then discusses six mooted responses, each of which have been explored to varying degrees in the last two years: blocking of social media services, deletion of stories, flagging of stories, changes to algorithms, bans to individuals, groups and communities, counterspeech and digital literacy. Finally section 4 considers the benefits of democratising Tech companies as a further part of resolving the dilemmas involved in pursuing some of these solutions.

2 THE ETHICS OF COUNTER-RADICALISATION

Radicalisation may be defined as 'the process whereby people turn to extremism' (Neumann, 2013). There is no one single process – people come to adopt extreme beliefs in a variety of different ways. Neumann identifies a tension in the concept of radicalisation between a focus on extreme beliefs and on extreme behaviour. A focus on extreme behaviour will usually refer to acts of violence and other serious illegality related to acts of violence, such as involvement in plots, planning or financing of them.

When counter-radicalisation strategy is focussed on violent acts, its justification closely corresponds to the justification of counter terrorism strategy. Preserving public safety and order is a preeminent duty of any state. In liberal states there is a strong presumption against interference with the individual's liberty or intrusion into their privacy, but the need to prevent serious crime and in particular acts that threaten life or serious injury is usually quite sufficient to outweigh this presumption.

There are of course deep ethical questions around whether radicalisation can itself legitimately be subject to criminalisation. Even in cases where somebody's speech clearly leads to another person's violent actions, this does not necessarily make them ethically responsible – individuals are responsible for their own actions. At most the radicalizers are indirectly responsible, perhaps in ways that are hard to register in criminal law. But even if acts of radicalisation themselves are not criminalised they might reasonably be subjected to official scrutiny and surveillance anyway – through the study of their communications with people reasonably suspected of carrying out or planning serious violence or injury.

What is it to hold extreme beliefs? Beliefs that are extreme when they depart far from received standards of reasonableness This response does not explain why holding extreme beliefs could be bad in any way that could merit the attention of law enforcement or intelligence services. Most 'extreme' beliefs in this sense are harmless. Indeed, many beliefs that were at one stage regarded as 'extreme' – such as a belief in the legitimacy of gay marriage, or women's suffrage – are now mainstream in liberal democracies. Furthermore, the 'extremists' who lead the way in arguing for these views are rightly regarded with gratitude, and applauded for leading public opinion in a more just direction.

An answer better equipped to justify official interest in radicalisation would emphasise extreme beliefs regarding the use of violence – say beliefs about the justifiability of killing or injuring civilians in pursuit of a political programme. This might appear to link extreme beliefs to actual violence. However, any sort of statistical link between the holding of these sorts of beliefs about violence and acts of violence is likely to be very weak.

Neumann accepts the inevitable contextuality of any assessment of 'extremism'. "Its content varies depending on what is seen as 'mainstream' in any given society, section of society or period of time" (Neumann, 2013, 876). But there are legitimate reasons for policing and intelligence services to scrutinise processes of radicalisation even when they do not lead to violence. For one thing it is difficult to disentangle processes of radicalisation that do lead to violence from those that don't. And furthermore, radicalisation not leading to violence may nevertheless "threaten democracy and social cohesion" (Neumann, 2013, 893). However, approaches to radicalisation that focus on extreme beliefs rather than violent behaviour "can be overly vague and distract governments' attention from the prevention of violence as their top priority. Most worryingly, they lend themselves to overreach, and—in the wrong hands—may be a licence for oppressing dissent" (Neumann, 2013, 893). To the extent that counter-radicalisation strategy does not target material with a demonstrable link to violence or serious harm, it will fall afoul of liberal objections. In what follows below this will be seen in relation to a range of mooted solutions proposed to the problem of Fake News.

3 THE RISE OF FAKE NEWS

In a recent discussion on Twitter, Jamie Bartlett, the former director of the DEMOS Centre for the Analysis of Social Media, makes a useful distinction between three kinds of Fake News:[1]

> There is monetised fabrication. People knowingly pumping out rubbish ('The Pope supports Trump') because it's a way of making money through Facebook / Google's ad-revenue

[1] https://twitter.com/jamiejbartlett/status/1061985678693670914?lang=en

> model...Then there is the knowingly false, misleading propaganda, pushed out for a specific political purpose. You all know we've always had this…Then there's the much bigger issue of highly partisan, one-sided, biased news. Sometimes unintentional, sometimes editorial. This is a huge grey area, and there's no technical or government or platform solution.

My focus on the second of these – knowingly false, misleading propaganda. As Bartlett notes, it is an old problem, supercharged by the speed with which social networking services have been able to spread it far and wide. Contemporary usage of the term 'Fake News' was largely established as a result of the 2016 US election. In the run up to polling day and aftermath of the result it was noted that large numbers of untrue stories – often made to resemble mainstream media resources like local newspapers – were spreading with enormous reach across social media. Above all, Facebook was identified as a key means by which these Fake News stories had spread.[2]

One notorious case is highly instructive for our purposes. This was the so called 'Pizzagate' case. This refers to a conspiracy theory which flourished across social media about a pizza restaurant, Comet Ping Pong. The allegation – thoroughly debunked – was that a paedophile ring was being run out of the Washington DC establishment, and servicing senior members of the Clinton campaign, including Bill and Hilary Clinton themselves. The case came to wider prominence on 1st December 2016 when Edgar Welch, armed with an AR-15 semiautomatic assault rifle, stormed the premises, apparently to save the children he sincerely expected to find there.

Fake News is also implicated in violence against the background of a number of deeply divided societies. In both Myanmar and Sri Lanka, false stories are alleged to have sparked intercommunal strife. For example, in Ampara in Sri Lanka, a false story spread about a Muslim plot to sterilise the local Buddhist population. A resulting altercation was filmed where a small restaurant owner, Farsith Abbam-Lebbe, appeared to confirm the rumour to a mob (as a Tamil speaking Muslim, not understanding the questions being asked in Sinhalese). The mob beat Abbam-Lebbe, destroyed his restaurant and burnt the local Mosque. The video of his 'confession' went viral.

Fake News may be implicated in violence as a direct spur – the attacks on Abbam-Lebbe, and Welch's raid on Comet Ping Pong seem to have

[2] See for example https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories or https://www.vox.com/new-money/2016/11/16/13637310/facebook-fake-news-explained

come as a result of encountering particular stories. Alternatively, Fake News may also be implicated at an earlier stage, in forming the world view according to which violence is the appropriate response to a particular story.

Throughout 2018 there has been a renewed focus on YouTube in particular as a source of online 'self-radicalisation' because of the huge amount of white nationalist content and conspiracy theories posted there. Some of the resulting coverage has cited cases like 'Pizzagate', where an identifiable threat to public safety is involved, but some stories have also a reflected a concern with promotion of extreme beliefs as a problem in itself. A number of stories highlighted the way YouTube's algorithm may push casual viewers to more extremist material. If one finishes watching one video on YouTube, it automatically loads another, making use of its profile of the user to play a video it thinks is likely to be interesting to that user. In practice this can take a viewer – over the space of just a few videos – to explicitly extremist content: racist, misogynist.

Conspiracy theories in particular thrive on YouTube. Conspiracy theories by definition are 'extreme' in Neumann's broad sense. A conspiracy theory posits a hidden set of causes of an action or event, but also an alternative view of how one may come by reliable information. It dismisses received confidence in the reliability of media sources, the government or science, usually attributing a deliberate motivation to deceive, 'covering up' the truth and preventing the public from knowing it. Quite why conspiracy theories should thrive so successfully on YouTube is not known. Part of the story will just be straightforward neglect: Fake News has flooded across a number of social media platforms, but most attention is restricted to Facebook and to a lesser extent Twitter, not YouTube. Tech companies have tended to react to problems like Fake News reactively, 'firefighting' as scandals arise. So far, most attention has been directed towards Facebook.

4 COUNTERING FAKE NEWS AS COUNTER-RADICALISATION

In what follows, I discuss some of the possible solutions which have been explored and discussed since the issue of fake news has come to public prominence.

4.1 Taking Down Social Media Platforms

Largely unthinkable in European Member States, blocking entire social media platforms has been employed as a temporary response to cases of Fake News, for example in Sri Lanka when it considered that there might be a spate of false stories like the one that led the mob to attack Abbam-Lebbe.

The potential for multiple threats to life arguably justifies taking down a social media service, if there are no better alternatives. Such a move, however, inevitably imposes high costs on large numbers of people by depriving them of a useful means of communication. Social media services both connect citizens with each other, and provide the possibility of having a platform to communicate with many at once. Suspending the service, even temporarily, will have an enormous impact. Furthermore, this moral cost may be higher in a society where there are fewer alternatives: if there are less developed alternative channels of media, and platforms on which citizens' voices might be heard. Nevertheless, the urgency of protecting life will still outweigh these considerations unless there are lower impact alternatives. If there are lower impact alternatives, such a measure will be disproportionate and unjust.

4.2 Removal of Stories

Removing Fake News posts from the social media platforms on which they have spread has been the solution most discussed in the media. It is also a solution actively being pursued, albeit it to a limited extent. But it is important to distinguish between different methods of take down, and the associated risks.

The first important distinction is between agents who might carry out the removal. State actors sometimes seek to use the law to mandate the removal of material by companies running the hosting platforms. Prior to the Fake News discussion, this might have arisen in relation to websites with illegal content. But there are two problems with this. First, a hosting platform may be in one jurisdiction and the audience in another. The information be legal to post in the platform host jurisdiction but not that of the audience. Second, there may be no clear legal definition of misinformation that is unlawful in either jurisdiction.

Furthermore, even if such an approach effectively limited the exposure of Fake news, it is not clear how that would advance the goals of counter-terrorism, especially where the content in question has no clear link to violence. The state seeking to limit expression by take-downs can easily look heavy handed and illiberal. Indeed, it can easily create more of an audience for the misinformation it has excluded.

Following the rise to prominence of issue in 2016, efforts to remove misinformation have mainly been discussed in terms of the obligation of platforms to remove false material. Facebook has been reluctant to delete stories merely for being demonstrably false, citing free speech concerns. Instead they have preferred the alternative of impeding their spread (as I discuss under the heading 'changes to the algorithm' below). Where they have accepted that material needs not only to be 'suppressed' but actually removed from the platform, that has been in cases of violence, like the attack on Abbam-Lebbe in Ampara discussed above. Facebook says its policy now is to remove material 'contributing to or exacerbating violence or physical harm'. On the basis of the policy, the stories about Muslims poisoning Buddhists have been deleted. The plan is for Facebook to work with local groups to identify the content that, relative to the local context, needs to be deleted. Facebook has repeated that it does not intend to take down all false material, preferring principles of free speech to win out. CEO Mark Zuckerberg chose the example of Holocaust Deniers – Holocaust Deniers falsely claim that the Nazis are innocent of genocide, but they are not thereby 'contributing to or exacerbating violence or physical harm'.

Finally, a state might attempt to remove content or block access to it without the assistance of the platform at all – that is to say deploying techniques which might be described as cyber warfare. Obviously such an approach raises very significant concerns and risks. Even more than using the law to mandate take-down, taking down without due process will be considered deeply disrespectful of norms of free speech and association. It would also be criticised as state interference with a private entity.

4.3 Flagging stories

An alternative approach taken by Facebook has been to flag stories verified as false, so that casual social media users simply stumbling across material can be alerted to its unreliability. A story verified as false and misleading ought to be singled out. Flagging a story as false may stigmatise a 'source', whether that is a media outlet or the individual sharing it on Facebook. In particular, declaring a story unreliable implies that those who have shared it are 'unreliable informants' themselves. But this might

be justifiable to prevent the spread of misinformation. However, Facebook abandoned this approach, publically reporting that their own research suggested that flagging stories as untrue could in fact result in users believing in the unreliable material all the more.[3] In its place Facebook has instead pursued a solution of linking alongside the article to more reliable reporting on the same issue.

4.4 Changes to Algorithm

Facebook resists banning individual stories from the platform, but it will 'demote' them, changing an algorithm to make it less likely that the story will turn up in a user's feed simply because friends are sharing it. Likewise, the basis for criticisms of YouTube in much of the coverage in 2018 concerned the way the algorithm for recommending new videos to watch promoted some and not others.

In January 2019 YouTube announced that it will recommend fewer videos that 'could misinform users in harmful ways'. This of course raises the question what kinds of harm they consider relevant. The videos of an anti-vaccination advocate can be very harmful, if it leads many parents not to vaccinate their children. And as we've seen, many argue that radicalising people by getting them to adopt extreme anti-democratic, racist or misogynist beliefs is harmful in itself.

YouTube's announcement is welcome, but it highlights the power of Tech companies and the opacity of much of their important decision making. This is an issue that is played out across a range of contemporary social issues. Algorithms increasingly determine more and more of contemporary life. O' Neil (2016) outlines some of the important matters where algorithms increasingly hold sway, such as policing and sentencing.

Algorithms are ever-present in our online lives through advertising. The reason the big tech platforms are free to users is that the more users a platforms has, the more useful data can be gathered and the more accurately adverts can be targeted at us. Developing algorithms to provide the kind of content most likely to hold our attention and keep us on the platform is arguably the central technological value these services offer. These algorithms are subject to no direct external scrutiny. It is only by observing their effect – spotting the fact that they serve up more and more extreme videos, for example – that they come to our attention.

Tech companies are understandably secretive about outsiders viewing what goes into making up the algorithm. It is central to their business

[3] See for example https://www.bbc.co.uk/news/technology-42438750

model and commercially sensitive intellectual property. Jamie Bartlett argues that the importance of algorithms to the public sphere means that we need to introduce outside oversight:

> Our law-makers – whether national or international – must create accountability officials who, like IRS or Ofsted inspectors, have the right to send in technicians with the requisite skills to examine Big Tech algorithms…This is especially true during elections, where governments must demand explanations and justifications for changes in news feeds and search results that might impact on the information the public receives.
>
> (Bartlett, 2018, 212-213)

But even if there were opportunity for outside scrutiny, it is questionable just how effective it could be. Carl Miller (2018) explains the powerful and somewhat arbitrary role algorithms increasingly play in our life in relation to a long conversation he had with an insider at a major Tech company.

> Within this tech giant, algorithms rarely stand alone. Instead, they exist within webs. 'I rely', he said, 'on signals that are produced by other algorithms.' His algorithm was fed by inputs that were shaped by other algorithms. It was like a car assembly line. He, like his colleagues, worked on a small, specific part of a much larger process...
>
> …The researcher scrolled through the bundle of instructions, and changed a single one to a two. A single value…'OK,' I said, 'what happened there?? Why did you change it? You know the two is wrong. But how do you know the one is right?'
>
> 'That', he said, gesticulating at the sabotaged result, 'is the point. It's a heuristic. I tried it, and it seemed to work. Then I tested it, and the result looked right. I can't say the one is true. I can only say that it passed minimum evaluation criteria. The whole algorithm is full of parameters that could have been something else. Truth is dead,' he sighed. 'There is only output.'
>
> 'Who checks there?' I asked.
>
> 'Me.'
>
> 'What about your boss?'
>
> 'You've seen how difficult it is to really understand. Sometimes I struggle with it, and I created it. The reality is that if the algorithms looks like it's doing the job it was supposed to do, and people

> aren't complaining, then there isn't much incentive to really comb through all those instructions and those layers of abstracted code to work out what is happening.' The preferences you see online – the news you read, the products you view, the adverts that appear – are all dependent on values that don't necessarily have to be what they are. They are not true, they've just passed minimum evaluative criteria.
>
> (Miller, 2018, 275-277)

This suggests a problem for the idea of external oversight. First, it is very difficult technically to understand what these algorithms do. I have written above that from the point of view of wider society we are only able to understand the potential ethical problems raised by dependence on algorithms by observing them as they arise. The extract above suggests that the position is not entirely different as far as companies' own internal processes go – problems are often identified after the fact, rather than being anticipated from the code itself.

4.5 Banning of Individuals or Communities from Social Media Platforms

Another solution that many platforms have favoured over actually banning content has been removing individuals, pages or forums from their platforms. This of course takes place for reasons that have nothing to do with Fake News – abuse or harassment are much more commonly the basis. But now pages set up simply to generate misinformation, such as money-making 'clickbait' sites, are subject to removal even if the news stories they share are not.

There is much scope for blurring of these distinctions – posts may simultaneously spread misinformation, be abusive or promote harassment. Furthermore, Tech Platforms' process for taking such decisions are also frequently opaque. When Facebook removed Alex Jones and his Infowars media company (at the same time he was banned from Apple, YouTube and Spotify), they cited glorification of violence and use of hate speech. Reddit, a platform which does not police false information or conspiracy theories at all, and which has a very robust norm of free speech, has nevertheless banned a page dedicated to discussion of the Pizzagate conspiracy theory. In practice this page repeatedly violated Reddit rules about posting personally identifying information about people entering the restaurant: on this community tantamount to an accusation of child abuse.

Platforms' (relative) comfort with banning communities and individuals reflects the smaller number of people consciously affected when people posting misinformation are removed. 'Banning' in this way is of course highly stigmatising as well – publically declaring the poster's speech as dishonest. But stigmatisation is sometimes morally appropriate. Indeed, the stigmatising effect is arguably epistemically beneficial in calling attention to a potential informant's unreliability. By the same token, of course, unfairly banning someone can count as a serious injustice, and platforms ought to be careful to have good reason to ban posters, and to provide means for contesting the decision.

Banning of discussion pages or forums is less problematic than banning particular people posting misinformation. No individual is silenced by taking away a specialised space for discussion of a niche topic: discussion can always continue on other forums. It takes away something that may be useful for participating in an online community, but it does not threaten rights in the same way as banning individuals does.

4.6 COUNTERSPEECH

Easiest to reconcile with liberal norms are acts of counterspeech. Counterspeech at its most basic is a matter of critically responding to posts, or being critical of its sources. Counterspeech might take the form of replying to an account sharing a Fake News story to say that it has been debunked, possibly linking to more reliable information to demonstrate this. Or it might take the form of a longer full story or video carrying out a debunking of prominent claims. Thus the work of fact checking sites such as Snopes.com or Politifact, can count as instances of counterspeech. So too can the work of debunking YouTube conspiracy theory videos, by making YouTube videos of their own in reply.

The extent to which there is any tension between counterspeech and liberal norms largely depends on who the agent engaging in counterspeech is. Citizens (or indeed non citizens) acting in a private capacity are entitled to their free speech, which straightforwardly includes the freedom to counter speech they disagree with. Counterspeech is more often talked about in relation to its possible role in countering trolling – hostile interaction online. At a workshop, the Horizon2020 Media4Sec project revealed dangers in counterspeech in the sense of directly responding to hate speech online:

> …police and civil society groups expressed deep ambivalence about the likely consequences of counterspeech. On the one hand, visible counterspeech can mitigate the harms of

> hatespeech and is often satisfying and reassuring to victims. However, it can also raise the level of trolling behaviour overall. This is because counterspeech is often met with an increase in trolling behaviour in response. Also, counterspeech easily descends into trolling itself. The conventional wisdom to not 'feed the trolls' is a recognition that what the troll usually wants is a response that can be further weaponised.
>
> (Roosendaal et al, 2018, 17)

Confrontations with purveyors of Fake News can easily degenerate into hostile encounters. One might worry about this because of the existence of sharers of Fake News online whose primary motivation is to provoke anger and increase tensions. But one might worry about the combustibility of confrontations online even between people who are entirely sincere. Increasingly encounters on media like Twitter tend to edge away from nuance and towards hostility. None of this is to deny the good that online conversation can do in steering people away from misinformation. It is merely to note that there is no guarantee that interventions will be effective.

Should states engage in counterspeech? To an extent it cannot avoid doing so. When false accusations of wrongdoing are made against a government and spread far on social media, it is appropriate that the government responds. Likewise, government may have a duty to respond to the spread of rumours and misinformation when they pose a threat to its citizens such that it cannot avoid making official statements on matters of public interest. However, for agents of the state to take part openly in online discussions can raise problems not raised by private citizens doing the same. It might be considered to have a chilling effect on free speech for state agents to intervene uninvited in discussions on Twitter or Facebook.

So while counterspeech may be a highly effective response to the threat of misinformation, it may be more effective the less it is perceived as a government sponsored activity. Indirect efforts, like policies that protect and promote a free and diverse quality press, for example, may be better. This of course is a non-trivial task currently, when newspapers are inevitably under competitive pressure from online platforms.

4.7 Digital Literacy

With the rise of widespread misinformation and conspiracy theories. there is greater discussion of digital literacy: the capacity of citizens to under-

stand the material they encounter online, and accurately evaluate its credibility. This is not simply a matter of education. The point is often made that a belief in certain conspiracy theories,thrives amongst people with a greater than average level of education. Such is the conclusion of David Aaronovitch writing in 2009:

> …conspiracy theories originate and are largely circulated among the educated and the middle class. The imagined model of an ignorant priest-ridden peasantry or proletariat, replacing religious and superstitious belief with equally far-fetched notions of how society works, turns out to be completely wrong. It has typically been the professors, the university students, the artists, the managers, the journalists and the civil servants who have concocted and disseminated the conspiracies.
>
> (Aaronovitch, 2010, 325)

This assessment predates the association of conspiracy theories and Fake News with social media. Nevertheless the point stands that education by itself is no safeguard against this kind of thinking. Aaronovitch has also described susceptibility to conspiracy theories as a kind of failure of historical understanding – not a failure to know some list of facts, but a failure to understand how social events do and do not happen:

> …fraught though the understanding of history is, and although there can be no science of historical probability, those who understand history develop an intuitive sense of likelihood and unlikelihood. This does not mean they are endorsing the status quo. As the great British historian Lewis Namier wrote, 'the crowing attainment of historical study is a historical sense – an intuitive understanding of how things do not happen'.
>
> (Aaronovitch, 2010, 7)

Similarly, Karl Popper opposes conspiracy theories and social scientific explanations generally in his Conjectures and Refutations. Attacking what he calls 'The Conspiracy Theory of Society' as a throwback to a primitive explanation of social phenomena, he presents the aim of the social sciences as explaining phenomena in terms of unintended consequences:

> I think that the people who approach the social sciences with a ready-made conspiracy theory thereby deny themselves the possibility of ever understanding what the task of the social sciences is, for they assume that we can explain practically everything in society by asking who wanted

> it, whereas the real task of the social sciences is to explain those things which nobody wants – such as, for example, a war, or a depression
>
> (Popper, 1972, 14)

These are comments about the difference between a deep social scientific understanding of social phenomena versus simpler, easy-to-understand narratives that lean on supposedly malign motivations of identified enemies. Most efforts at Digital Literacy concern the acquisition of beliefs in the first place rather than questions of how beliefs might be fit together into a coherent world view.

There are a number of efforts currently aimed at improving the capacity of people – particularly young people – to distinguish between reliable and unreliable information encountered online. For example one current scheme in Belgium schools --Lie Detectors' --send journalists into schools to introduce children to techniques for verifying material, and how to carry out basic credibility checks (as well as introducing them to the principle that they should not simply trust material online but rather need to verify it first). Another similar scheme in France introduces children to a conspiracy theory video the scheme has constructed itself (exposing that the CIA was responsible for spreading AIDS – a conspiracy theory from yesteryear), and uses that a basis for practically demonstrating how to detect that material is fraudulent (as well as showing how quickly a video like that can spread online and be promoted by people ideologically attracted to its conclusions.

Greater digital literacy on the part of the general population, if it can be achieved, is a solution without ethical costs – it is simply beneficial to equip people with better skills at distinguishing reliable from unreliable information. Acquiring this ability may compete with acquiring other goods, including other educational goods (there are only so many hours in the school day), but unlike many of the other solutions we have discussed, it does not incur moral risks. However, it is no solution at all in the immediate term – a better educated citizenry is a long term solution. As things stand, counter-radicalisation strategy faces the reality of a population now with very mixed levels of digital literacy, and so will continue to face the dilemmas of more morally costly measures.

5 COUNTER-RADICALISATION, PRIVATE COMPANIES AND DEMOCRACY

Several of the solutions mentioned above in part 3 pose dilemmas for the proper boundaries of freedom of speech and the role of the state in the public conversation. And the answers are highly significant for the kind of

society we are going to live in for the coming decades. However, they are decisions that are essentially being made by private entities. Big Tech companies are private businesses, yet they play host to the public conversation whereby people come to decide what to believe. The monopoly position of the Big Tech companies has brought vast profits, but increasingly Big Tech companies find themselves falling short of public expectations in a way which directly affects these profits. Big Tech is increasingly subject to what has been referred to as a 'techlash', as public trust in these entities drops dramatically. Miller (2019) among others, argues that this is best explained by their enormous power over increasing areas of our lives, combined with a lack of democratic control or input. Furthermore, they now govern increasingly large proportions of our public space: spaces where we pursue not only leisure and social interaction, but much of what now makes up civil society. These private companies thus play a central role in our democracies.

Earlier we discussed deletion of news stories, banning of individuals and algorithms determining which posts are seen or videos automatically loaded. To the extent that these actions affect public spaces they not only have negative consequences for the individual but adversely affect their rights of expression. If these companies continue to provide sites of civil society interaction, it is undesirable for the character of these interactions to be left only to the tech companies. Assuming responsibility for these questions also leaves Big Tech with the problem of answering these dilemmas alone. A better alternative would be to find ways to increase public participation in determining how free speech might operate on these platforms.

What models of public involvement are currently available? Miller mentions user involvement models such as Reddit, Wikipedia and the Internet Governance Forum. Reddit is a very interesting example of open and accountable community moderation, and also a test site for the hypothesis that targeting extreme acts can be pursued consistent with highly permissive norms of tolerance for extreme views. This mirrors the distinction of part 1 between the two approaches to counter-radicalisation, radicalisation resulting in violent acts versus one that would widen focus to extreme views as well. Reddit has pursued a strategy of only pursuing/deleting/banning material that has a direct adverse effect on others: –

> 'Do not post content that encourages, glorifies, incites, or calls for violence or physical harm against an individual or a group of people; likewise, do not post content that glorifies or encourages the abuse of animals.'

However, even within this relatively narrow definition of unacceptable behaviour, the company itself has to make choices about how to ban communities not upholding these standards. Furthermore, Reddit is criticised precisely for permitting forums promoting extreme ideas to flourish, some explicitly accusing it of hosting ‘radicalised communities’

Likewise, Wikipedia provides a model for how community participation can result in a resource that – so far – has resisted attempts to corrupt it for political purposes and maintained its accuracy. Unlike Reddit and social media, Wikipedia has a narrow mission of providing reliable information. Social platforms like Facebook, Twitter and Reddit all seek to provide a space in which a much wider conversation can take place. Nevertheless, both Wikipedia and Reddit demonstrate the power of community editing. Hosted by the UN, the Internet Governance Forum brings together representatives from an enormous range of stakeholders: governments, companies, academia and NGOs. It has no power, but is a useful site for discussion, and represents one of the few real world examples of accountability by multistakeholderism. A more detailed and practical model for introducing public accountability and transparency by involving stakeholders in the setting of policy is possible. The primary questions for such an approach would be ‘Which stakeholders are relevant to innovation by these companies?’ and ‘How ought this stakeholder interaction to take place?’ The Reddit/Wikipedia models of community editing offer a bottom-up model for increased accountability/transparency. Multistakeholderism offers something that could be described a horizontal model. Both could be an advance the current patterns of decision by private companies and would be likely to introduce a much wider range of approaches to structuring our civic online spaces.

REFERENCES

Aaronovitch, David. Voodoo Histories. London: Vintage

Bartlett, Jamie. 2018. The People VS Tech. London: Ebury Press

Kudlacek, Dominic, Laura Treskow, Brendan Marsh, Stephanie Fleischer, Matthew Phelps and Maja Halilovic Pastuvic. Pericles Deliverable D1.2 Gap Analysis on Counter-Radicalisation Measures available at http://project-pericles.eu/wp-content/uploads/2017/10/Pericles-D1.2-Gap-Analysis-Report.pdf

Miller, Carl. 2018 The Death of the Gods. London: William Heinemann

Neumann, Peter. 2017 'Countering Violent Extremism and Radicalisation that Lead to Terrorism: Ideas, Recommendations, and Good Practices from the OSCE Region'. At http://icsr.info/wp-content/uploads/2017/09/Countering-Violent-Extremism-and-Radicalisation-that-Lead-to-Terrorism-2.pdf

Neumann, Peter. 2013 'The trouble with radicalization' in International Affairs vol 89 no 4 p 873–893

O' Neil. Cathy 2016. Weapons of Maths Destruction: How Big Data Increases Inequality and Threatens Democracy. USA: Penguin Random House

Roosendaal, Arnold, Kat Hadjimatheou, John Guelke and Petra Vermeulen. Medi@4Sec Deliverable D4.5 Workshop 5: Policing of Trolling on Social Media: Ethical and Legal Issues available at http://media4sec.eu/downloads/d4-5.pdf

Popper, Karl. 1972. "The Conspiracy Theory of Society" in David Coady ed. 2006. Conspiracy Theories: The Philosophical Debate. Aldershot: Ashby Publishing