Workshop 5: Policing of Trolling on Social Media
Ethical and Legal Issues
Deliverable 4.5
University of Warwick and TNO

Co-funded by the Horizon 2020 Framework Programme of the European Union
The research leading to these results has received funding from the European Union’s Horizon 2020 Research and Innovation Programme, under Grant Agreement no 700281.
Contents

Executive Summary ................................................................. i
1. Introduction.................................................................................. 3
   1.1 MEDI@4SEC ........................................................................... 3
   1.2 Work Package 4 ....................................................................... 3
   1.3 Deliverable 4.5 ......................................................................... 3
2. Introduction to ethical and legal issues with policing of trolling on social media........... 5
   2.1 Distinction between ethics and law ........................................... 5
   2.2 Summary of crosscutting ethical and legal issues ....................... 6
3. Ethical and Legal Issues with policing of trolling on social media and measures to prevent, disrupt, and enforce ................................................................. 8
   3.1 Pathways to prevention, disruption and enforcement .................. 8
   3.2 Prevention .............................................................................. 10
   3.3 Disruption ............................................................................... 13
   3.4 Enforcement: Law enforcement monitoring of social media content ......................................................................................................................... 18
4. Conclusions................................................................................... 19
   4.1 Summary list of Recommendations ............................................ 19
Executive Summary

Trolling refers to a very broad category of online nastiness, much of which is not illegal and is easiest dealt with by private means such as moderation implemented by social media platforms themselves.

Key legal and ethical issues

The prosecution of trolling faces distinct challenges, chief amongst which is the lack of a mature legal framework; insufficient training and skills amongst LEAs and prosecutors; the lack of a harmonized international framework that would enable cross-border cooperation; and the unwillingness of victims to pursue prosecution once offensive material has been successfully taken down.

The most common legal issues arising in connection with anti-trolling activities concern data protection, privacy, and freedom of speech of those who post trolling material and people who might be mistakenly labelled as doing so.

Gaming is a key site of cyberbullying and hatespeech exposing very young children to harm online but has been largely overlooked by government, civil society groups, and schools.

While identifying users of social media would be useful for police trying to enforce anti-trolling laws, regulations requiring mandatory identification of users of social media are illegitimate from a legal perspective, as they infringe upon privacy in a disproportionate manner and lack a legal basis.

Blocking content (takedown of trolling material) satisfies victims, silences trolls, and improves the environment online, but it also leaves trolls unpunished and potentially undeterred. It can also limit freedom of speech.

Automatic moderation and removal of content by social media providers can be cheaper, faster, and impose less of a psychological burden on people than human moderation. But human moderation is likely to be more nuanced, accountable, and respectful of freedom of speech.

Trolls weaponise anti-trolling activities to satisfy their desire to provoke offense and cause conflict. Counter-speech in particular can be manipulated by trolls in ways that lead it to descend into offensiveness and hate.

Coordinated action, involving a range of actors from civil society to industry, government and law enforcement is essential for trolling of any kind to be effectively prevented, countered, and punished.

Summary list of key recommendations

1. Technical solutions ought to be pursued by specialist companies and social media platforms. But they need to be pursued with the understanding that their assessments (e.g. of what is hate speech or just free expression of political speech) will sometimes be disputable, and provide recourse for individuals who wish to contest e.g. a decision to block them.
2. Employers whose staff are put at risk of trolling as a result of the work they do have a duty of care to support those employees and protect them from harassment and abuse. A wider conversation is needed about the extent of their obligations. The following is an indicative list of actions employers might take:
   - Train their employees in protecting themselves from trolling (NGOs could provide courses)
   - Provide support in the form of strategies for responding to trolling
   - Consider purchasing technological solutions to protect staff from trolling
   - Provide counselling and support
   - Assist with take downs of behalf of their employees
   - Assist with pursuit of prosecutions on behalf of their employees

3. Social media platforms should consider greatly expanding the role of human moderation, as this is promising as a means of making social media environments safer, but this will require more resources, both in terms of numbers of moderators, but also in terms of support for them.

4. Central and local governments should fund greater training for teachers, parents and children in schools. There are NGOs in a position to provide this training, but the model needs to be expanded.

5. Governments should fund and academics should pursue research into the utility and pitfalls of counterspeech. This kind of research is important for those seeking to provide well-founded advice to those who approach them for help and through training provided to schools, companies etc.

6. Governments should better define the legal responsibilities of social media providers and gaming sites with regard to trolling. Most current measures relate to unlawful content, but do not necessarily take into account individual interests of (potential) victims.

7. Governments should try and formulate a clear legal definition of trolling in order to facilitate enforcement.
1. Introduction

1.1 MEDI@4SEC

MEDI@4SEC focuses upon enhancing understanding of the opportunities, challenges and ethical consideration of social media use for public security: the good, the bad and the ugly. The good comprises using social media for problem solving, fighting crime, decreasing fear of crime and increasing the quality of life. The bad is the increase of digitised criminality and terrorism with new phenomena emerging through the use of social media. The ugly comprises the grey areas where trolling, cyberbullying, threats, or live video-sharing of tactical security operations are phenomena to deal with during incidents. Making use of the possibilities that social media offer, including smart ‘work-arounds’ is key, while respecting privacy, legislation, and ethics. This changing situation raises a series of challenges and possibilities for public security planners. MEDI@4SEC will explore this through a series of communication and dissemination activities that engage extensively with a range of end-users to better understand the usage of social media for security activities. MEDI@4SEC will seek a better understanding of how social media can, and how social media cannot be used for public security purposes and highlight ethical, legal and data-protection-related issues and implications. Activities centre around six relevant themes: DIY Policing; Everyday security; Riots and mass gatherings: The Dark Web; Trolling; and Innovative market solutions. MEDI@4SEC will feed into, support and influence changes in policy-making and policy implementation in public security that can be used by end-users to improve their decision making. By structuring our understanding of the impact of social media on public security approaches in a user-friendly way MEDI@4SEC will provide an evidence-base and roadmap for better policymaking including: best practice reports; a catalogue of social media technologies; recommendations for EU standards; future training options; and, ethical awareness raising.

1.2 Work Package 4

The role of Work Package 4 (WP4) is to identify ethical and legal issues arising from the use of social media for public security and policing purposes as well as to facilitate ethics procedures within the project. Ethics and legal issues will be reported for each of themed workshops.

1.3 Deliverable 4.5

This report presents ethical and legal issues around the policing of trolling that were raised during the MEDI@4SEC workshop on Trolling held in London on 15th May 2018. In total, 50 people participated in the workshop, of which 30 were external (not part of the MEDI@4SEC project). Participants came from France, Denmark, Portugal, Germany, Greece, Italy, the Netherlands, Romania, Slovenia, Austria, Switzerland, Sweden, and the UK. Non Medi@4Sec participants included police officers (12), local and central government (5), NGOs (4), the tech industry (4) and research organisations (5). The morning sessions were devoted to presentations designed to provoke and stimulate
‘Trolling’ refers to a very broad category of online nastiness. Trolling can be usefully defined as the targeting of defamatory and antagonistic messages towards users of social media. Trolling broadly understood includes: cyberbullying; cyberhate; cyberstalking; cyberharassment; revenge porn; sextortion; digital vigilantism or ‘digilantism’; naming and shaming; and flaming. Cyberhate is trolling targeted towards those with minority status. The term ‘flaming’ is used by different actors to refer to different activities: some use it to refer to extremely provocative language, designed to start a fight; others use it to refer to those who post offensive or provocative material for their own entertainment or gratification. All the activities listed here involve the targeting of individuals for abuse online. Revenge porn, exploitation and doxing involve specifically exposing aspects of someone’s private life online and thereby ruining their reputation. Most of these acts take place within a context of wider abuse, including offline, but some trolls, in particular those who belong to the sociologically distinct group of individuals who self-identify as trolls, carry out the vast majority of their abuse online. The legal status of all the trolling-related acts just listed differs from jurisdiction to jurisdiction and from act to act.

In this report, three approaches towards trolling are discussed: prevention, disruption, and enforcement. For each of these, a distinction is made between machine mediated responses and human responses. Machine mediated responses are all types of responses working autonomously based on algorithms or Artificial Intelligence (AI) or human direction. These include automated takedowns, blocking or disruption of trolls and offensive material; changes to settings to enable user-controlled access and exposure to material or accounts pre-assessed as trolling. Human responses are those responses that involve direct human interaction with material, trolls, or legal frameworks. They include changes to legal and policy frameworks; online interventions such as counter-speech and the use of human moderators to decide material what to take down; support for victims and those vulnerable to trolling; education and training for children, parents, teachers and employees working in vulnerable roles.

2. Introduction to ethical and legal issues with policing of trolling on social media

2.1 Distinction between ethics and law

The ethical and legal issues discussed here are distinct but related. Ethics relates to the reasons we have for thinking something is the right or wrong thing to do. Law relates to the rules a society has in place with which people can be coerced legitimately by the state into complying. Some ethical issues are reflected in the law; but ethics is broader than law: while illegal behaviour is typically also unethical, much unethical behaviour is not regulated by the law. For example, ethics tells us that breaking a promise is nearly always wrongful, even if it is sometimes justified overall given the costs of keeping the promise in the particular circumstances. Thus, ethical considerations always come into play when promise-making and promise-breaking are concerned. In contrast, legal considerations only arise when the promises in question fall within the narrower set that are legally enforceable, say in the form of contracts. Ethics is thus broader than the law, providing perspectives from which to criticize the law and argue for its reform.

To say that something is legally permissible is to say that it can be done without legal consequence. Thus legal analysis can tell us what the law requires and so what we need to do if we want to avoid being prosecuted, sued, and so on. Ethics cannot do this. But because the law usually reflects some minimum threshold of acceptability, the fact that something is lawful does not by itself demonstrate that it is desirable – let alone morally laudable or an example of best practice. Ethics can offer guidance as to what we are permitted to do, encouraged to do and required to do, morally speaking. Thus it can offer advice about what kinds of things public security providers and technology developers should do if they want their actions to be both legal and examples of best practice. Many police codes of ethics and codes of conduct include both minimum standards of behaviour and more aspirational principles. In practice, there is significant crossover between ethics and the law. For example, Privacy by Design (an approach to projects, including technology design, that promotes privacy and data protection compliance from the start) is an approach that began as best practice but is now being incorporated into EU Data Protection law. And ethics can have a key role to play in informing the law when, as is so often the case, the latter lags behind technological development.

A crossover between ethics and law also arises when both sets of principles, rules or norms share common roots. The legal issues raised in this report are grounded in and make reference to international, EU, and domestic legal instruments. The ethical issues are grounded in the values of liberal democratic political theory, including in particular equal treatment, fairness, autonomy and liberty. Because the legal instruments referred to in this report are those of progressive liberal democracies, the legal and ethical

---


5 Seminal philosophers in this field include: John Locke; Immanuel Kant; John Rawls; Ronald Dworkin; J.S. Mill.
concerns described below are closely related and sometimes overlap. For this reason we have presented the legal and ethical issues together in Section 3.

Sometimes the values underpinning ethics and the law conflict, and ethical and legal debates are often concerned with finding the best way to reconcile conflicting duties and obligations. Within liberal democratic societies the question of which rights, rules and requirements these values give rise to is also often subject to debate. Finally, the interpretation of the role and status of the police and the nature and source of their authority varies significantly between liberal democratic countries. As this suggests, shared roots and basic commitments do not preclude dilemmas and disagreements within both ethics and the law.

2.2 Summary of crosscutting ethical and legal issues

From an ethical perspective, where possible prevention of trolling activity can be achieved this is a more desirable outcome than either successful enforcement or disruption. It is an ethically better outcome for offences not to take place at all than for them to take place and be effectively prosecuted. Preventive measures also raise few legal issues because they do not often entail interference with anyone’s rights. Rather, they involve activities such as education and training.

Where prevention is not possible, disruption is often the next best option from an ethical point of view. It is an intervention that interferes in real time with the offence to either prevent its completion or mitigate its effect. Takedowns, blocking, hacking back and throttling are all examples of disruptive activity that thwarts the intentions and plans of trolls. Disruption has many benefits including technological simplicity, speed, relatively low cost, improving the online environment for all by filtering out toxic elements, and satisfying victims of trolling. However, it also raises more legal and ethical issues than prevention, because it silences people or prevents them from communicating and so potentially interferes with their freedom of speech or their ability to use important communication platforms. Those subject to disruptive measures seldom have the chance to challenge them.

Automated disruptive measures are quick, cheap, and relatively simple compared to, for example, the use of human moderators. They also spare or save moderators the psychological and other costs of continued exposure to trolling and other abusive material online. However, human moderators are able to make far more nuanced judgements of the offensiveness of material than automated alternatives. And they are relatively easier to hold accountable.

While counter-trolling actions by online communities on social media and social media platforms themselves are generally more effective than the actions of public authorities, they cannot substitute the strong arm of the law in cases of serious harassment, stalking, and abuse of individuals online.

But the enforcement of anti-trolling laws faces a number of challenges, chief amongst which is the persistent lack of legal clarity about what is worth prosecuting and what is not. Enforcement is rendered easier if anonymity is denied to users of social media, but
anonymity is valuable in other ways. For example, it is important to people's ability to exercise freedom of speech and to enjoy a right to privacy online.

All of these issues are complicated by a lack of agreement on where moral/corporate/legal responsibility for responding to trolling lies. Social media platforms and gaming platforms are able to act far quicker with less bureaucracy and legal constraints than police, but there is ongoing debate about the extent of their responsibility to do so and the extent to which they do so in a way that is transparent and accountable. If the policing of the online sphere is to be left primarily to these actors then much closer and more transparent cooperation between them and law enforcement, civil society and government should be promoted.
3. Ethical and Legal Issues with policing of trolling on social media and measures to prevent, disrupt, and enforce.

In the sections below we will discuss potential approaches to meeting some of the challenges posed by trolling. First, we identify crosscutting legal pathways to supporting better anti-trolling activities across all kinds of intervention. Then, we consider potential solutions at each stage (prevention, disruption, enforcement) in more detail.

3.1 Pathways to prevention, disruption and enforcement

3.1.1 Clarify distinction between legal and illegal trolling

From a legal perspective, one of the key challenges posed by trolling is that laws currently used to enforce norms against abuse, harassment and hate online suffer from a lack of clarity. This makes it difficult to determine what crosses the line of the law, and what doesn’t. Key elements to most anti-trolling laws are: deliberately causing serious offense or distress through communication, as well as harassment. However, from a legal perspective, determining what is a harmless prank and what is harassing or cyberstalking is open to interpretation. To a large extent Law Enforcement Agencies (LEAs) are having to judge for themselves what is worth prosecuting.

It is therefore unsurprising that workshop participants mentioned that the “creating of a framework, of what is allowed online, and what is not” is key to help LEAs in finding, addressing and prosecuting trolls.

Along the workshop, several times the call was made for establishing an EU-wide legal framework defining offensive content and the competences LEAs have with regard to such content. Such a framework may be difficult to achieve, though, as what types of content are flagged as ‘offensive’ may be culturally diverse along the different EU Member States.

3.1.2 Clarify legal and moral responsibilities of different actors: LEAs; social media platforms and gaming sites; employers of those trolled because of their work; schools.

A range of different actors play important roles in preventing, disrupting and enforcing rules against trolling in all its forms. Yet there is lively debate about where the boundaries of each of their responsibilities lie. LEAs have a remit that stops with activity that is criminal under the law, but the lack of a mature legal framework means that the line between legality and illegality is still largely being drawn-by police and prosecutors themselves.

Social media platforms and gaming sites provide a forum for trolls, and this raises questions about their role in cleaning up their sites, even of material that does not meet the threshold of illegality. The workshop saw wide criticism of social media companies who were argued to evade their responsibilities in many cases. Some of this has been a matter of resisting the extension of various kinds of regulation to extend to their spaces as well. One example is resisting having the status of a content provider. If Facebook or Twitter were treated as content providers they would be subjected to a large range of
A further question around moral and legal responsibility for trolling and its effects arises in relation to employers of people whose work exposes them to abuse, harassment and hate online. Staff working in high-risk occupations like politics, journalism, social work, teaching and policing are all particularly likely to encounter trolling in the course of their normal duties. For example the International Women’s Media Foundation and the International News Safety Forum reports that 88% of women journalists feel at increased risk because of social media. There is a case to say that in professions like this employers have a duty of care to protect staff from what is increasingly an occupational hazard. Where this is an inevitable and unavoidable part of the role (it would be wrong for a journalist to stop investigating an issue because Internet trolls objected), employers may be obliged to provide counselling, or training on how best to respond to trolling activity. Or as Magnus Boyd at Schilling Law firm argues, they may be obliged also to help with getting material taken down.

Education is a key part of the picture across multiple stages of development. Arguably the duty of care on the part of employers of workers likely to be at risk extends to training employees to effectively respond to this threat. And as children are increasingly at risk as well, education in the risk needs to become standard in school classrooms. There are a number of initiatives by NGOs in a position to provide this kind of specialist training, including for example by Romania’s Safer Internet Centre. Police are also in a position to assist here because of their growing experience working with a range of victims.

---


9 Twentyman, J. 2015. ‘How best to protect your employees from Internet trolls’ Financial Times July 14th https://www.ft.com/content/fdea7aa9-262d-11e5-bd83-71cb60e8f08c accessed 29th of June 2018

3.2 Prevention

3.2.1 Restricting anonymity: mandatory registration of users of social media to enable identification of trolls

Some Law Enforcement Agents in the workshop wished for an automatic requirement to have members of social media platforms identify themselves in a certified way before opening an account. This is commonly discussed as a solution to trolling in public debate about how to tackle it. Different social media platforms have different levels of anonymity as it is, with a number dispensing with it altogether. It was suggested in the workshop that national governments could make rulings to force all social media companies to do without anonymity. This would make it easier to identify the users who act as trolls and to prevent them from creating new accounts using fake identities. A qualitatively good and trustworthy method for online identification would be required to achieve this.

3.2.2 Challenges to restricting anonymity

Even though there are quite some initiatives in online identification and authentication, it may be difficult to find a common international standard for identification on the web. There are European requirements and levels of accountability as defined in the eIDAS regulation. Nevertheless, implementing these on a platform that is available in several EU countries (and outside the EU) may be quite demanding for the platform providers, in particular if these are small and open platforms. Depending on the system and the business model of an authentication provider, the costs for each login may come at the expense of the platform provider. So, the more the platform is used, the higher the costs, which works countereffective for free, open platforms (sometimes with idealistic intentions). A less difficult option for preventing anonymous posts that came up during the workshop is to prohibit VPN connections on social media platforms. A VPN (Virtual Private Network) arranges a connection by a virtual deviation: by connecting to another server, the IP address (and thus possibly the country of origin) changes. VPNs offer three things: Security (data is encrypted against parties who want to mine data such as hackers and governments), Anonymity (your IP-address, a personal ID-code, is changed to the IP of the VPN server - you can no longer be recognized), and Freedom (through a VPN you are able to connect to servers over the entire world – online censorship is no longer an issue). VPNs are therefore gaining in popularity. However, from the LEAs perspective, it is hard to figure out who is actually behind the IP address – which is difficult if you want to find and block trolling.

3.2.3 Negative consequences of ending anonymity: freedom of speech and privacy

And there are limitations to the impact ending anonymity would have. After all, services like Facebook’s which dispense with anonymity still have problems with trolling and offensive content. While some trolls are likely emboldened and enabled by the perception that they can abuse and harass without ever having to account for themselves, it was noted in the meeting that some trolls are sincere and 'righteous', perceiving themselves
to be harassing others for legitimate reasons and would be happy to continue to do so even if they were forced to stand by what they had written without anonymity.¹¹

Requiring identification beforehand also implies that anonymous posts would be made impossible. This might also result in hindering the freedom of speech. For instance, when individuals want to post sensitive content or disclose certain information that would require the protection of the source. An example is the Dutch political cartoonist Khalid Guédar, drawing satirical pictures of the Arabic world. He received death threats after re-posting a caricature made by the Jordanian writer-journalist Nahed Hattar, who was shot to death on his way to the court of Amman (Jordan). Guédar was shocked by the threats, as "the purpose of placing that drawing on social media was to provide information and pay tribute to a man who gave his life for the freedom of drawing. (...) I notice that signs in the Islamic world are constantly under pressure. I did not expect these threats at all, because you see the drawing everywhere on the internet. Besides, I did not add my own comment, except for a quote from draftsman Charb." However surprised Guédar might have been – it is imaginable that with a filter on sensitive content, the post of Guédar would have been blocked up front, because the content could be offensive to people from the Arabic world. This could limit the freedom of speech on social media.

From the above options, two main issues raise from a legal perspective. One is related to privacy, if anonymous posting of content is made impossible. The other is a limitation of the freedom of speech. Both the right to privacy and the right to freedom of speech are fundamental rights. These are now examined in detail.

### 3.2.4 Limitations on the right to privacy

A mandatory requirement to identify oneself when making use of a platform means that anonymous communication is limited. There are legitimate reasons to want to be anonymous in discussion online, linked to the right to privacy. For example, some people may benefit from being able to discuss and receive advice on embarrassing medical conditions, or aspects of their personal life or religious beliefs, anonymously.

Apart from the risk of people not feeling safe to contribute or to share content, the need to share personal information in order to be able to participate on a platform infringes upon privacy. However, in order to decide upon the relevance of this, it is necessary to determine whether this limitation is legitimate. In this case, users do not have a choice whether to share their personal data if they want to use the platform. But of course, they do have a choice whether or not to use the platform. Moreover, requiring identification would apply to all users, while the aim is to block or monitor only a few users acting as trolls. For ID-checks, privacy issues come up. First of all, the protection of privacy is considered a fundamental right under article 8(1) of the Charter on Fundamental Rights of the European Union and article 16(1) of the Treaty on the Functioning of the European Union, which state that everyone has a right to protection of personal data concerning him or her. Furthermore, the new General Data Protection Regulation (GDPR) sets out

---

¹¹This conclusion is also supported by Rost et al. 2016, Digital Social Norm Enforcement: Online Firestorms in Social Media PLoS ONE 11(6)
privacy safeguards. In Article 5, general principles are listed, and in Articles 16-22 rights of data subject are set out. In particular, these rights have been strengthened and the responsibility of data controllers to demonstrate their compliance with the GDPR is made the central element of this legal instrument. For a limitation on the right to privacy to be legitimate, a couple of requirements have to be met. First, there has to be a basis in the law. Second, the limitation has to be necessary in a democratic state, and, third, the limitation has to meet the requirements of proportionality. The three requirements will be briefly elaborated upon below.

The first requirement is the legal basis. There are laws that prohibit certain content or constitute defamatory content as unlawful. However, there is no legal provision determining that access to open platforms requires identification of the users. During the workshops, several comments came up relating to the need for a common legal framework on trolling and legal means to facilitate enforcement. For prevention purposes, however, the identification requirement might remain problematic.

The second requirement, that the measure has to be necessary in a democratic state, can be argued for. Basically, the call for the measure of mandatory identification follows from the need to prevent trolling in order to avoid people becoming the victim of trolls. So, certain measures can be necessary to protect the rights and freedoms of citizens.

The third requirement is that the infringement needs to be proportionate in relation to the aim of the measure, and there have to be no less infringing means available that can lead to a similar effect (subsidiarity). As indicated, requiring identification of all users may be disproportionate in relation to the aim of only blocking or monitoring a few of them. There are often alternative means to identify individual users if they behave illegally or post unlawful content. Basically, in these cases LEAs are provided with means to look up IP addresses, subscriptions, and other ways to track down an individual. These means become available as soon as there is a concrete trigger, such as an unlawful act or illegal behaviour. This also implies that the requirement of subsidiarity has not been met.

Based on the above, it can be assumed that requiring mandatory identification of all users, instantiated by LEAs, would constitute an illegitimate limitation of the right to privacy. Creating a basis in the law does not solve this issue, since the measure is disproportionate and the aim of the measure can most likely be achieved by less infringing means as well.

To conclude, the expectation is that, based on the current legal frameworks, mandatory identification of all users is a disproportionate infringement on the right to privacy. So, government related actors, such as legislators or LEAs cannot require this.

3.2.5 Limitations on freedom of speech

A similar test a for the infringement upon privacy is applicable to the limitation of the freedom of speech. Anonymity allows people to benefit from being able to discuss taboo moral and political issues in a way they might be reluctant to in the open, especially if they live in a society in which there is limited toleration of political dissent. So again, there has to be a basis in the law, the infringement has to be necessary and proportionate in a democratic state, and it has to relate to the purposes mentioned in Article 23 a) to f) GDPR.
In this case, differences may occur in what is deemed unlawful content or hate speech and what is not. There is no EU-wide common legal framework or defined set of terminology to qualify content as trolling. However, some types of content can be unlawful or illegal and therewith be blocked. The necessity in a democratic state is probably more difficult here, as it depends highly on cultural attitudes to what is acceptable and when LEAs need to get involved or the extent to which (potential) victims have to take care for themselves. In particular, with regard to content and freedom of speech difficulties arise, since the more democratic a state is, the more the freedom of speech is respected.

3.3 Disruption

3.3.1 User controlled an openess/exposure to potential trolling

Besides the option of blocking content from being posted, an alternative is to limit user access to the content. So, content can be posted, even if it is recognized and flagged as ‘offensive’. Accessing the content, however, might require an opt-in from the users. Without the explicit opt-in, content flagged as ‘offensive’ will not be seen.

This kind of solution does not interfere with freedom of speech and it empowers users to decide for themselves what they want and do not want to be exposed to. It does not impose one interpretation of offensiveness on users of social media. There is a risk that the more sophisticated user-defined content provision becomes, the more we will retreat into ‘echo-chambers’ in which we are not exposed to provocative or diverse ideas. This is often warned against as encouraging the fragmentation of civil society.

Another possible user-centric anti-trolling measure is the sharing of user-defined blocklists. This was discussed as a bespoke product by one of the speakers. Sharing of blocklists may also have ethical costs as well as benefits. By multiplying the effects of a block, anything that makes the initial blocking of an individual questionable becomes more significant. There will doubtlessly be borderline cases where a person is blocked in all sincerity by a victim who perceives them as trolling when in reality their conduct amounts to legitimate criticism or reasonable anger. More worryingly trolls themselves may seek to weaponise any anti-trolling measure as a further means for harassment of victims – i.e. exploit it as a tool for further trolling. A common trolling tactic, it was pointed out in the meeting, is to attempt to get victims banned themselves, by submitting claims that they have broken rules of conduct themselves. Some commentators complain that these kinds of rules end up amounting to a ‘hecklers veto’. If the effect of unfair bans is intensified this may rise to the level of injustice.

3.3.2 Taking down trolling material and blocking users

Taking down offensive material is quicker, easier and cheaper than employing counter-speech or attempting to identify and prosecute trolls. Taking down offensive material not only protects direct victims, but it also protects other social media users from exposure to it and generally helps to sanitise an often toxic online environment. It is the priority of the victim, and it also satisfies them. More than 50% of victims of illegal trolling whose

crimes are reported to the Metropolitan Police's Online Hate Crime Hub abandon any attempt to press charges once the offending material in question is taken offline. This is a weighty consideration in their favour, although there is an argument to say that potential future victims, or even norms of justice, are better served by prosecutions.

Police involvement with take downs is limited to passing relevant cases on to civil society groups that then petition tech companies on behalf of victims.

The widely, publically discussed ethical cost raised by taking down of material (as well as by prosecution discussed further under ‘enforcement’) is whether this can represent censorship, and/or pose risks of the so called ‘chilling effect’. The chilling effect takes place when people feel a disincentive to attract attention, becoming more reluctant to speak in a particular forum or on a particular topic. It was argued by UK police that the present threshold for action being taken on material is so high that chill cannot be argued to obtain. However, it can be responded that since harms like chill relate to public perception, that it is not enough that actions in practice respect freedom of speech but that this be effectively communicated as well. An open and liberal public domain requires institutions that are trusted as well as trustworthy

Takedown and blocking also raise a legal concern, as they potentially constitute a limitation on the freedom of speech. Without clear definitions of ‘offensive content’ it is difficult or impossible to automatically take materials down or collect evidence. Basically, content should be compared to certain characteristics or a database of comparable content. And, automatically taking down content based on algorithms can lead to situations of automated decision making with a legal effect, since the target is limited in his right to freedom of speech. Moreover, if selection takes place based on characteristics of the user (the alleged troll), it also includes the processing of personal data. Processing personal data in order to take automated decisions with a legal effect or otherwise comparably impact on the individual is essentially prohibited.

Depending on the way the blocking of content is implemented, the measure can be proportionate. If only content from certain users or specific types of content is blocked, the impact is limited to the content that is deemed to be of an unlawful or illegal nature. There is no blocking in a general sense, and people who behave properly are not limited in their options to share content.

So, it can be concluded that there is no general answer as to whether freedom of speech is legitimately limited by blocking certain content. However, if content is clearly unlawful or defined as unlawful, the measure of blocking content may be necessary and proportionate.

---


14 As laid down in Article 10 (Freedom of expression) of the European Convention for the Protection of Human Rights and Fundamental Freedoms.

15 A legal effect is a decision with any form of legal significance, such as imprisonment, a fine, or the ability to exercise a right.
3.3.3 Human vs automated moderation for takedowns and blocks

Another key real time intervention is moderation of digital social spaces. This can be effective – for example a number of Reddit boards are effectively self-policed by enthusiastic volunteers as a service to their community. Reddit is able to depend on large numbers of working hours on the part of people motivated to facilitate discussion on a topic of their interest. Different environments are able to have very different standards for what counts as acceptable behaviour, where lighter or heavier moderation would be appropriate. In the case of large companies like Facebook moderation has a less ambitious aim: only excluding the very worst, most abusive behaviour. This is difficult, nuanced work, often carried out under pressure, and largely involves looking at unpleasant behaviour in order to determine whether it violates the content policy. There are significant concerns for the welfare of moderators, some of whom have complained to the press about working conditions and being forced to view extreme material, such as child pornography.

Some companies now offer automated moderation. Technical solutions like this offer increased sophistication over automatic filtering discussed under prevention, and could avoid some of the problems of blacklisting terms. Blacklisting affects individuals using terms innocently, whereas moderation ought to be a less blunt tool. In the same way that a human being can read extra information into the context of use, artificial intelligence is increasingly able to as well. However, human viewers remain much better than machines at understanding much. Furthermore, where the impact of banning someone is a significant loss, decisions made by humans are likely to be more legitimate. The future is likely to involve artificial intelligence tools assisting human moderation in many cases. The example of Reddit speaks to the usefulness of moderation for enabling a better quality of online human interaction, but also demonstrates that it is intensive work. Moderation could have a large impact on trolling in many contexts, but only if it becomes normalised that this is nuanced work and requires paying for many working hours in order that it is done properly. Furthermore, if it is work that puts its employees at risk – exposing them to extreme, disturbing material – the employer’s duty of care will oblige them to take according action, whether in the form of training, counselling or other support.

3.3.4 Using tech to fight back: ‘hacking back’ and ‘throttling’

**Hacking back**

An option to disrupt trolling that goes a step further is the hacking of accounts and computers of trolls. Hacking as such is prohibited and constitutes a criminal act in several

---

16 For example see this researchers post about her research on moderation processes: Robert, S. ‘Behind the Screen: commercial content moderation’ [https://illusionofvolition.com/behind-the-screen/] accessed 29th of June 2018.


18 See for example: [https://www.smartmoderation.com/] accessed 29th of June 2018

19 See for example [https://netino.com/5-reasons-not-to-automate-all-your-moderation/] accessed 29th of June 2018.
countries. It is therefore not an easy way of disrupting without breaking the law. Nevertheless, in some countries, such as the Netherlands, there is a debate going on, including around new legislation, around the legality and ethics of allowing the police to hack back criminals if they have hacked a computer network, or to gain access to one’s computer via other devices in a network.

While it would be morally impermissible for law enforcement to hack accounts used by trolls to monitor or interfere with their activities, companies who run the platforms on which trolls carry out their offences are less ethically constrained, and indeed arguably have a responsibility to monitor and control what is possible to do on their service. Increasingly companies are attempting to do this – for example it has become common for Twitter to ban prominent extremists or provocateurs20. Although the policies of these companies are subject to heated arguments over whether standards are correct or monitoring services effective enough, as is the case with Facebook’s recently revealed moderation guidelines21.

**Throttling**

Where trolling activity is identified one tactic that was discussed as a very effective response is so-called ‘throttling’ – reducing the perpetrator’s connection to the service and thereby greatly reducing their ability to misbehave on the platform. There are a number of features which make this an effective response. The perpetrator cannot be sure it is happening, as there are always a range of other technical explanations for their reduced access to the service. There is no forum to contest their treatment, and the move sidesteps any need for companies to provide guidelines for acceptable behaviour which can be further weaponised by trolls. However, all these features make the measure ethically questionable as well. There is not a recognised ‘human right’ to use these different platforms, but there is an increasing recognition that access to many of them is very important to participation in everyday modern life.

It is easy to imagine measures like ‘throttling’ being directed against service users in error, with correction very difficult because of the lack of transparency. A key argument for the moral seriousness of trolling is that it excludes victims from digital social spaces. The principle that there is a default moral entitlement to participation online is undercut by an endorsement of arbitrarily withholding access in a way that is not open to challenge. The opacity of such a measure also reduces the possibility of a person being given a chance to amend their behaviour in response to the threat of reduced service. Measures like this will be more or less objectionable on the basis of the site – some sites like Facebook are much more plausibly claimed to be a central part of everyday social

---


interaction. But even a ‘fringe’ space, say one dedicated to a quite esoteric hobby, could potentially become quite an important part of a user’s life.

From a legal perspective, issues around freedom of speech might arise from throttling. Throttling implies that trolls are being recognized based on either their accounts or on the content they are posting. Recognition can, thus, be based on two main approaches: identification, which can be in the form of accounts or in the form of mandatory identification to be able to create an account as described above or scanning of content before it is posted on the platform.

The scanning of content as such may not have legal consequences. In addition, delaying the content of being posted is not a strict form of censorship as well. Trolls are still able to execute their right to freedom of speech, albeit in a slightly delayed manner. For the scanning, defining the specific purpose and what types of content is scanned for has to be determined beforehand though. This is necessary to prevent the scanning from becoming a general scanning measure looking for several types of content other than trolling as well. Basically, for the processing of personal data, which would be the case here, the aim of the processing and what data are processed have to be specified beforehand. A too open formulation would imply that there is a lack of purpose specification, rendering the processing illegitimate. For instance, it could become interesting to track behaviour and networks of users on social media platforms, or to look for copyrighted materials or other forms of unlawful or illegal content. In particular, if this would be done by LEAs, the data collected would qualify as police information, implying that specific legal regimes on police information handling would become applicable.

3.3.5 Counterspeech

A final key intervention discussed was so-called counterspeech. Counterspeech is publicly and visibly contesting hatespeech. Counterspeech avoids objections from free-speech advocates because it doesn’t silence anyone, rather just adds further speech onto the record. However, 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. But some groups are finding innovative ways of disrupting this vicious circle. For example, the Berlin-based Peng! Collective uses bot technology used by spammers to detect misogynistic language on Twitter and then to counter-troll it with humorous messages in an attempt to disrupt trolls.

The patterns of counterproductive interventions can be seen in, for example, the Institute for Strategic Dialogue manual on how to do counterspeech. This poses law enforcement

---

22 Supporting this point, Research by DEMOS and Facebook report examples of counter hate-speech initiatives that have been effective at overshadowing and/or drowning out hate speech.

23 Institute for Strategic Dialogue manual on how to do counterspeech; Positive Messages Project: Heartmob
and anti-trolling groups with a range of further dilemmas – how should online hostility be treated when it is a response to the online hostility of others? What is the appropriate response in cases where people are both victim and abuser? For example, if aggression is met with an aggressive response, are police obliged to treat all abuse equally?

3.4 Enforcement: Law enforcement monitoring of social media content

One solution mentioned in the workshop was setting up a EU database with recurrent images, so that Law Enforcement Agencies can analyse problematic content easily, and sharing of data within EU borders is easier. Another mentioned solution was cooperation with Social Media providers – for instance to collect data in order to facilitate prosecution. While both could be profitable, legal issues are likely to arise. As described in deliverable 4.4, acquiring, processing and sharing data in such a way is subject to legitimate grounds:

Personal data “shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes (‘purpose limitation’);” (GDPR, Article 5)

As deliverable 4.4 states, this means that the processing (meaning any action, including saving, copying, viewing, deleting, etc.) of personal data has to be based on a legitimate ground as provided in Article 5 & 6 of the GDPR. And that there has to be a specific purpose defined beforehand. Generally monitoring social media is not specific enough and does not naturally relate to a purpose, so it is needed to determine what the ultimate aim is. A purpose in terms of generally monitoring social media to prevent crime or to contribute to public security is not specific enough, since it cannot be clear what data exactly can be expected to be processed in which cases. A clearer description is required, such as a specification of the types of crime (e.g. hate speech or defamation), in combination with an indication of the data that will be processed and under which circumstances.
4. Conclusions

This report has presented ethical and legal issues arising in connection with anti-trolling activities, discussed at the MEDI@4SEC Trolling workshop. Key issues discussed included the need for better defined responsibilities of different anti-trolling actors and more clarity and international consistency on what kinds of trolling online are illegal. Law enforcement has an important but limited role to play in the fight against trolling, because trolling is ultimately an issue anchored in the culture of human interactions. As this implies, education is a key preventive anti-trolling measure, and there was strong consensus in the meeting that this should be proactively rolled out in schools, including for very young children. Education should be age-specific taking into account the very different concerns and attitudes of children at different stages of development and should go beyond the current focus on staying safe online to discuss the harms of trolling and what it means to be respectful online. Finally, the volume of trolling is such that debates about restricting the freedom of speech of those identified as trolls online seem to be losing their force. Without action against trolls, the internet risks becoming such a toxic environment that freedom of speech is de facto being chilled as people either opt out of social media or stop taking part in important public debates, leaving them to be colonized by trolls.

4.1 Summary list of Recommendations

1. Technical solutions ought to be pursued by specialist companies and social media platforms. But they need to be pursued with the understanding that their assessments (e.g. of what is hate speech or just free expression of political speech) will sometimes be disputable, and provide recourse for individuals who wish to contest e.g. a decision to block them.

2. Employers whose staff are put at risk of trolling as a result of the work they do have a duty of care to support those employees and protect them from harassment and abuse. A wider conversation is needed about the extent of their obligations. The following is an indicative list of actions employers might take:
   o Train their employees in protecting themselves from trolling (NGOs could provide courses)
   o Provide support in the form of strategies for responding to trolling
   o Consider purchasing technological solutions to protect staff from trolling
   o Provide counselling and support
   o Assist with take downs of behalf of their employees
   o Assist with pursuit of prosecutions on behalf of their employees

3. Social media platforms should consider greatly expanding the role of human moderation, as this is promising as a means of making social media environments safer, but this will require more resources, both in terms of numbers of moderators, but also in terms of support for them.

4. Central and local governments should fund greater training for teachers, parents and children in schools. There are NGOs in a position to provide this training, but the model needs to be expanded.
5. Governments should fund and academics should pursue research into the utility and pitfalls of counterspeech. This kind of research is important for those seeking to provide well-founded advice to those who approach them for help and through training provided to schools, companies etc.

6. Governments should better define the legal responsibilities of social media providers and gaming sites with regard to trolling. Most current measures relate to unlawful content, but do not necessarily take into account individual interests of (potential) victims.

7. Governments should try and formulate a clear legal definition of trolling in order to facilitate enforcement.