Rights, Equality and Citizenship (REC)

Programme of the European Commission

(2014-2020)

Monitoring and Detecting Online Hate Speech

4.3: Mandola WS4 Workshop with stakeholders

Abstract: The MANDOLA project hosted a Workshop about Challenges and Responses to Online Hate Speech on 15 December 2016 in Brussels.

The workshop invited participation from persons and organisations working in the area of online hate speech and encouraged participation from Law Enforcement, Internet Industry and Civil Society. There was active participation from the Mandola partner countries including Bulgaria, Cyprus, Ireland, France and Greece and others from countries outside this group. 25 persons from a range of stakeholders were invited to participate and contribute to the discussions by sharing their knowledge, expertise and experiences in this complex area.

Contractual Date of Delivery	October 2016
Actual Date of Delivery	December 2016 (Document: August 2017)
Deliverable Security Class	Public
Editor	Cormac Callanan
Contributors	All *MANDOLA* partners
Quality Assurance	Marios Dikaiakos

The *MANDOLA* consortium consists of:

FORTH	Coordinator	Greece
ACONITE	Principal Contractor	Ireland
ICITA	Principal Contractor	Bulgaria
INTHEMIS	Principal Contractor	France
UAM	Principal Contractor	Spain
UCY	Principal Contractor	Cyprus
UM1	Principal Contractor	France

Document Revisions & Quality Assurance

Internal Reviewers

1. Marios Dikaiakos (UCY)

Revisions

Version	Date	By	Overview
1.1	21/8/2017	Cormac Callanan, Aconite, Editor	Edits & Updates
1.0	17/7/2017	Marios Dikaiakos	QA
1.0	10/3/2017	Speakers & Presenters	Edits & Updates
1.0	1/3/2017	Cormac Callanan, Aconite, Editor	First draft of minutes.
1.0	15/12/2016	Cormac Callanan, Aconite, Editor	Event Hosted

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

The MANDOLA project hosted a Workshop about Challenges and Responses to Online Hate Speech on 15 December 2016 in Brussels.

The workshop invited participation from persons and organisations working in the area of online hate speech and encouraged participation from Law Enforcement, Internet Industry and Civil Society. There was active participation from the Mandola partner countries including Bulgaria, Cyprus, Ireland, France and Greece and others from countries outside this group. 25 persons from a range of stakeholders were invited to participate and contribute to the discussions by sharing their knowledge, expertise and experiences in this complex area.

Selected Quotes from participants

"If you cannot measure it you cannot improve it. We lawyers say 'if you cannot define it then you cannot protect it'. Hate speech is not always illegal. Not all speech is illegal. What definition is used for online Hate Speech?

"Speech is not about words. Speech is about interactions."

"this is the project, to explain the current capabilities and weaknesses."

"I understand now why Facebook and twitter find it such a problem deciding what hate speech is."

"Whatever you remove from the Internet well most of it has the potential to get back. The internet is the great big recycler"

"…brings the 'online' inline with human rights"

"It is impossible to tell the total amount of hate speech on the Internet. Those days are over."

"We are not talking here about what the law says. We are talking here about owning a company and whether allowing or not allowing certain material on your platform. It is a business decision. "

"If you need a definition of hate speech you just need to look at the criminal code. Hate speech is made by threats. Libel that is a crime. Hate speech in specific related topics that are crimes in local legislation."

"So one of the main problems when talking about hate speech is trying to find a higher level, a meta definition instead of looking at what the single criminal code says"

"We have magistrates to decide what a crime is under local law and what is not. This should be the focus because as soon as we abide by the law, the court decision is the only binding text we should be subject to."

"To motivate social media if there is something that is horrible that they don't want to remove then shout very loud and if you shout with 100k users and they will all of a sudden say that it is against their terms of service so let's remove it"

"In all parts of everyone's life, if we see something illegal happening. We think that the law should be enforced. Here, no."

"One of the surprising things we discovered is that members of the majority who are interested in promoting the rights of the minorities are also very frequent targets of hate speech and very often even against them is the most aggressive forms of hate speech"

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

"The most frequent propagators of hate speech? We asked 'From whom have you heard hate speech' and they are the ordinary people by far – 71%."

"almost 1/3 of the people who do not know that incitement to hatred is a crime and what is striking is that this number is much higher among the people who are supposed to be protected by this clause in our justice law"

"some of the harms suffered by victims of hate speech are the same of those experienced by people with post-traumatic stress disorder"

"some of the damaging psycho-emotional affects are a sense of anger which is one of the most common responses"

"fear can take on paranoid qualities and drastically disrupt the lives of some victims"

"There is a loss in faith in law enforcement and the whole criminal justice system"

"we went through all the legal texts that could be remotely be relevant to answering the question 'What is hate speech?'."

"these types of semantical debates are very common and if you are a lawyer you enjoy them but if you are a programmer, you kind of have this nightmare"

"it is a question of legal requirement engineering. The combination of these heatmaps plus what I read in the legal text is simply a stunningly difficult problem."

"Coding such definitions is by no means easy from a legal perspective."

"If I want negative sentiment I can buy that online as companies provide this "

"On the other hand you have to teach people that any filthy thought you have in your mind, you don't need to share it with the world!"

"We have to educate, educate, educate, If we do not start now in grammar school, then in 10 years we will say that old people don't get it and it will be too late."

"You can't effectively defend people that are infringing on other people's freedom of speech if the legal threshold for access to subscriber data is too high."

MANDOLA: Monitoring ANd Detecting OnLine hAte speech is a 2-year transnational project funded by the Rights, Equality and Citizenship (REC) Programme of the European Commission.

The project aims at improving the public understanding of the prevalence and spread of on-line hate speech and empowering ordinary citizens to monitor and report hate speech. Its objectives are:

- to monitor the spread and penetration of on-line hate-related speech in EU and its Member States using big-data approach, while investigating the possibility to distinguish between the potentially illegal hate-related speech and non-illegal hate-related speech;
- to provide policy makers with information that can be used to promote policies for mitigating the spread of on-line hate speech;
- to provide ordinary citizens with useful tools that can help them deal with on-line hate speech irrespective of whether they are bystanders or victims;
- to transfer best practices among Member States;
- to set up a reporting infrastructure that will connect concerned citizens with Law Enforcement and appropriate abuse desks and will enable to report illegal hate-related speech.

The project is implemented by the Foundation for Research and Technology - Hellas, Greece with the support of the partners: Aconite Internet Solutions (Ireland), the International Cyber Investigation Training Academy (Bulgaria), Inthemis (France), the Autonomous University of Madrid (Spain), the University of Cyprus (Cyprus) and the University of Montpellier (France).

Further information: www.mandola-project.eu

Contact: Prof. Evangelos Markatos,
Foundation for Research and Technology Hellas, Greece
Email: markatos@ics.forth.gr

JUSTICE
Rights, Equality and Citizenship Programme 2014-2020

Workshop Contributions

Cormac Callanan, Workshop Moderator

The ideal situation is that the legislation of the country is reflected in the commercial terms of business that you have with these organisations and that they mirror each other in a way that is understandable by everybody.

We have not dealt with the issue of prevention comprehensively because that is not our focus in the Mandola project and we have done very little research in the area of counter-narrative. This is about responding to online speech by engaging in counter-argument and showing some ideas about how alternative thinking can work. This is a different area of debate not covered in this workshop.

It was unexpected that the legal landscape would be as complex entanglement as it is. In the area online child sexual abuse material many years were spent encouraging governments to agree a common definition which was almost 99% the same in many cases. We have nothing like that in the online hate speech area.

Contributions to the workshop were very impressive demonstrating a wide range of wisdom, expertise and practical experience. It was very interesting to learn how many practitioners approach the problem of illegal online content and specifically online hate speech. The legal issues were very provocative. It was very enlightening to learn how practitioners approach and overcome the obstacles of online hate speech.

Session 1
Current approaches to online hate speech

Mandola Project Aims, Objectives, Progress Achieved by Mr. Evangelos Markatos, FORTH

All the partners of this project are active in the area of cybersecurity. Some of the partners are active in their national Centres of cybercrime. The partners have different roles in the project. What is monitored? What techniques are used? In this project basically

Mandola wants to monitor the spread and penetration of online hate speech in the EU. Lord Kelvin once said "*If we do not measure it we cannot improve it*." Measurements are at the core of solid conclusions. If you do not have good measurements, you do not have good conclusions, you cannot make good decisions.

In the first pass, Mandola applies a hate detection filter using the HATEBASE database which is basically a database of hate related words. In parallel with this first filter, sentiment analysis is performed to remove false positives. Real humans (raters) are then asked to annotate the negative tweets and comments that have been occurred from the previous step, whether are hate or not. Also, the raters are asked to assign the tweets/comments to one (or more than one) of the following categories of hate-speech: ethnicity, nationality, religion, gender, sexual, disability, class. The output of this work will be used to train and validate the classification algorithm of Mandola dashboard.

Mandola generates a HeatMap which basically displays hateful related content as red on the map. If there is a lot of red, then there is a lot of hate. If there is blue then there is very little.

Mandola uses the twitter API feed because in addition to being open it includes the geographical coordinates. If someone makes a hateful comment it can be traced it to a geolocation. Traffic on twitter can suddenly pick up and a negative comment causes others to just jump in.

The International Network Against Cyber Hate (INACH) - Research – Report – Remove
Ms. Suzette Bronkhorst

INACH addresses all forms of discrimination and members believe that INACH adds value to the internet. INACH brings the 'online' inline with human rights and unites national, international and local organisations aswell as individuals who are active in the implementation of human rights on the internet.

INACH promotes online respect, responsibility and citizenship through countering cyber hate extremism and violent incitement and raising awareness of online discrimination. INACH actively enforces human rights and mutual respect for the rights and reputations of all internet users securing a safer internet. INACH is active in many areas.

Cyber hate is a problem that is huge, growing and a danger to our societies especially since social media started. There is a big difference between illegal hate speech and wider hateful speech which directly affects people. This needs to be defined. Commercial internet companies do have definitions but they don't seem to understand what their definitions are.

JUSTICE
Rights, Equality and Citizenship Programme 2014-2020

The International Network Against Cyber Hate (INACH) - Research – Report – Remove

Mr. Ronald Eissens

INACH is currently implementing a project funded by the European Commission called **Research, Report, Remove** countering the Cyber-Hate phenomena. INACH will produce a greater structure on Cyber-Hate and counter strategies. The project will produce recommendations for the social media industry and is updating its manual on how to recognise Cyber-Hate. The heart of the project will be a database on Cyber-Hate fed with online complaints from all over Europe.

In May 2016, the EC signed a Code of Conduct on countering Hate Speech online with Facebook, Twitter, Youtube and Microsoft. The code of conduct requires signatories to review within 24 hours all illegal hate which is brought to their attention from organisations and individuals.

The EC asked the INACH network to assess the level of compliance with the code of conduct during 6 weeks. This was achieved by issuing reports to the companies which in the opinion of INACH was illegal material. INACH also issues other reports to them which are not necessarily illegal according to the framework decision or the national penal code but are not acceptable to the social media companies themselves according to their own terms of service. A number of INACH members and other organisations produced 600 cases.

The verdict was not very good for the social media services due to non-compliance in most of the cases of the material not being removed within 24 hours and non-compliance on material that was clearly illegal but they did not remove it. They wanted legal opinion.

INACH has existed since 2002 but why does INACH do it? In the short term, the clean-up is good. Policing is always necessary. You do it also because of the victims. Freedom of speech is important but some speech is bad for society and/or leads to violence in the short or long term.

The latest trends in hate speech indicate a shift from the classical Nazi and racist sites to the main stream and it all goes into social media – the two big areas of concern are anti-Semitism and hate against Muslims. Right now hate against refugees is quite 'topical'. The violent component of hate speech is also going up.

It is impossible to tell the total amount of hate speech on the Internet. Those days are over. In the 90's INACH could give fairly good estimates then in the 2000's INACH gave guestimates. Now INACH can just say… it is BIG…Well it is.

Nothing is perfect. This is the future for INACH. There is no silver bullet to get rid of hate on the net. There is a combination therapy. There is a lot of knowing what is going on. Registration and mapping needs to continue. Having a good complaints system and having a good database is important. Using trusted reporters works to a certain extent but there are always problematic disagreements. Counter speech is very important but is very hard to do since it is very labour intensive.

European Digital Rights (EDRI)

Mr. Joe McNamee

EDRI is a collection of NGOs who work on defending fundamental rights in the digital environment.

The first EDRI heard of the Mandola project was the report on the implementation of the framework decision. The commission is congratulated for going to the effort for funding this research which is very necessary and it shows the legal challenges.

Commissioner Jourova stated last week said that 28% reports led to deletion and this could be improved. Why? It could be 100%? Would that be perfect? Imagine if we could delete anything we want from the Internet, simply by reporting it. INACH's could be deleted if we wanted. Mandola's could be deleted if we wanted. All one has to do is submit a report and create uncertainty for the service provider and they will remove it. Brilliant!??

No, EDRI disagrees. EDRI believes that the challenge relates to predictability. Either as a society we believe in democracy and we believe in law or we don't. EDRI believes in law. EDRI believe in democracy and law because it creates possibilities for accountability, redress and review. If you don't have law you have arbitrary treatment of fundamental rights and, once things start going bad in society, the people that suffer most from arbitrary decisions are the people that are weakest in society.

It is not a coincidence that every relevant piece of international law says that restrictions on rights need to be provided for by law. In the European charter of fundamental rights restrictions have to be provided for by law they have to respect the essence of the rights and freedoms in the fundamental rights charter. They must be

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

proportionate and they must only be made if necessary and genuinely meet the needs of general interests recognised by the Union. Restricting peoples fundamental freedoms is a last resort and has to be done in an accountable way.

The first section of the EC Code of Conduct is all about illegal material, fighting illegal content, comprising of a range of principles.

In the second, operative, part it is clear that in relation to hate speech, in relation to incitement to violence, it is the companies that are taking the lead. It appears to EDRI that we have abandoned the notion of law and we have abandoned the notion of accountability. It goes on to say that the internal rules of the companies will encompass the law and the companies will check the reports against their terms of service first and, where necessary, the law. Well, if the law is part of their terms of service it means that it will never be "necessary" to check against the law.

Arbitrary restrictions on fundamental rights are unacceptable.

Legislation needs to define the role of industry because the human rights challenges of the whole initiative and the role of the code of conduct or the role of terms and conditions of a business are a danger if we allow it to be the only way to address societal problems. The terms of service of these companies are deliberately unclear. They want to be able to act arbitrarily. The Code of Conduct allows them to act arbitrarily and it encourages them to act arbitrarily. No other stakeholder has any responsibility in the code of conduct apart from the companies and the obligations of the companies are not exactly clear.

I thank the Mandola project for its initial research. I think it is very good. I think it is great that there is detailed analysis of the legal definitions, because some of us like to live in a democratic society. It is important that reliable statistics are being produced in order to allow more efficient policy making, accountable public policy making.

Read more: https://edri.org/faq-code-conduct-illegal-hate-speech/

Session 2
Legal and Psychological Aspects of Online Hate Speech

Open Society Institute - Sofia, Bulgaria
Ms. Ivanka Ivanova, Law Program Director

The Open Society Institute - Sofia is a non-profit organisation and serves as the operator for the NGO programme of the EEA Grants in Bulgaria. The major resources of the NGO Programme are allocated to promote democracy, equality and human rights, to fight racism and xenophobia in Bulgaria and OSI needs to ensure that these funds are well spent.

OSI wanted to have a baseline. OSI wanted to know what the situation is like now. After 5 years of supporting projects of NGO's, OSI wanted to know if OSI had changed anything. OSI also wanted to know in what specific areas, in what specific policy measures there is the biggest need to intervene. So, OSI conducted a nationally representative public opinion poll.

OSI wanted to study the incidence of hate speech but it was quite clear in the beginning that the words hate speech itself cannot be used because of the lack of appropriate and widely accepted Bulgarian language translation of the concept. The survey asked people "Have you encountered in the public space statements which expressed hatred, aggression or disapproval towards minorities". This definition per se already puts some limits because it only measures what people hear, it only measures speech, not other forms of expression.

One of the surprising things discovered is that members of the majority who are interested in promoting the rights of the minorities are also very frequent targets of hate speech and very often even against them is the most aggressive forms of hate speech.

Hate speech is a relatively widespread phenomenon in Bulgarian society. In 2013, 2014 and 2016 about half the people responded to say that they have heard expressions that contain aggression or disapproval or hatred towards minorities and in 2016 there was a visible increase in the number of people who report having heard such incidents. In the first two years it was about 47% and in 2016 it was already 58%.

In Bulgaria, Roma are, by far, the most frequent target of hate speech. Turks were the second most frequent target of hate speech and in third place were gay people. In the first three years it was quite obvious that we had to deal with the negative stereotyping of these minorities – Roma, Turks and gay people. The autumn of 2013 though, was the time when the migrant influx towards Europe increased. Many of the immigrants passed through Bulgaria which coincided with a radical increase in hate speech against Muslims.

It was interesting to investigate whether the different minorities are victims of different negative stereotyping which should inform

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

different approaches of the different NGO's that want to promote tolerance towards a specific minority. The survey tested what were the most common associations with the word "criminal" and in the first two studies, the largest majority said that they did not associate any of the groups put in front of them with the word "criminal" but in the 2016 study there is already a visible decline in the people who said "none of these". Roma are very frequently associated with "criminal". Immigrants are also associated with "criminal" but to a much lesser degree. Minorities such as gay or Jews are almost never associated with the term "criminal" at all which means that if someone wants to promote tolerance towards Roma, in my simple view, has to first enlighten others that they are not criminals and this has to be the focus of the counter speech. If society wants to promote tolerance against gay and towards refugees and immigrants and Jews, the same approach would not be appropriate. They have to look at the negative stereotypes which undermine the public image of the respective minority.

More than 70% of the people who have encountered hate speech have encountered it on television. Shops and bars or other places of social communication are the second most frequent place. Public transport is also quite an important place of encountering hate speech. Internet is currently sharing second place as the most popular place to encounter hate speech but it should be noted that it is probably much less important than the TV since in Bulgaria only 44% of people say that they use the internet every day. The survey asked "*From whom have you heard hate speech*" and "*the ordinary people*" are mentioned as propagators of hate speech by 71% of the respondents. Politicians and journalists are identified frequently as hate speech propagators.

In terms of strategies to fight hate speech, probably the best it is to target the small groups who are not supposed to use hate speech at all, public servants, business men or so called experts are identified by 10% of respondents as propagators of hate speech.

The survey asked if the respondents support criminal prosecution of hate speech; a large majority of people said "yes" but when we asked "are you likely to report to the authorities hate speech" only 23% said "yes". Almost 1/3 of the people do not know at all that incitement to hatred is a crime and what is striking is that this number is much higher among the people who are supposed to be protected by this clause in the criminal law. A huge disincentive for reporting hate speech to the authorities is the fear that the witnesses or the victims are going to lose their job. We have registered a number of cases where they say "*we have heard it*", "*I hear it all the time from my boss in the working place and I am afraid that I will lose my job*".

An important problem related to the design of the survey related to the question we ask people about the incidence of the hate speech expression itself, about the label "hate speech". The survey had to test three different words because actually the English language expression is rather telling but there are many languages in Europe in which you have to translate it descriptively which already makes the whole research endeavour rather problematic.

Cyprus Neuroscience and Technology Institute
Ms. Aliki Economidou

CNTI is a non-profit, non-governmental organisation employing 34 individuals on a full-time basis, with many interns and part-timers. CNTI has projects focusing on the future orientation of human brain technologies and social transformation, as well as on humanity related issues. CNTI aims at building inter-linked socio-techno-cultural works through science and dialogue, having as vision to explore and utilise the evolution of information and communications technologies for strengthening the process of peace building and civil education. CNTI has three units – the New Media Lab, the Global Education Unit and the Humanitarian Affairs Unit.

The New Media Lab (NML) is active in both research and social intervention. It has experimental projects that aim to develop new theories of learning, new IT-rich and mobile-based curricula and technologies to assess the role of emotions, attention and mental states in learning. The New Media Lab has developed a diagnostic tool known as MAPS - Mental Attributes Profiling System; a language independent screening test, relying on cognitive rather than language-based measures, capable of predicting children at risk (possible dyslexics) and equipping teachers with a profile of their mental abilities so as to design personalized remediation programs. .

The Global Education Unit promotes and supports active global citizenship through local, European and global initiatives. Its projects focus on equipping youth and educators with knowledge, skills and tools to increase awareness about global issues and to encourage action for a more just and equitable world; engage citizens of all ages in

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

intercultural dialogue and support peace-building initiatives. The Humanitarian Affairs Unit is a major department at the organization and directly responds to the needs of the vulnerable population. The unit implements a UNHCR funded project, providing legal aid and social support to asylum seekers and refugees.

The Cyprus safer Internet Centre (CyberEthics) has been operating since 2006 by the New Media Lab and has been funded by the Safer Internet Programme of the European Commission. It operates a Helpline, a Hotline and an Awareness Node. CNTI is a member of the INSAFE and INHOPE networks. The Cyprus Cybercrime Centre of Excellence (3CE) is an additional project of the NML and operates under the three pillars – training, research and education.

CNTI is interested in the psychological aspects of hate speech; some of the harms suffered by victims of hate speech are the same of those experienced by people with post-traumatic stress disorder such as, panic, fear, anxiety, nightmares, intimidation and denigration. Some of the damaging psycho-emotional affects are a sense of anger which is one of the most common responses to being the victim of hate crime, arising from a deep sense of personal hurt and betrayal. Victims experience feelings of powerlessness, isolation, sadness and become suspicious. Their fear can take on paranoid qualities and drastically disrupt their lives, such as not leaving home on frequent basis. The individuals might stop using public transportation where they might be victims of hate speech and lose faith in law enforcement and the whole criminal justice system. It has been also observed that the victims of hate speech are apprehensive in reporting cases of hate speech as they hold the belief that there is nothing that can be done. Most victims report changes in their lifestyle such as in the way they walk, they answer the phone and talk to strangers.

There is an interesting quote worth recalling because it shows how a victim of hate speech feels and how hate speech scars the victims far more deeply, "*You are beaten or hurt because of who you are. It is a direct and deliberate and focused crime and it is a violation of a person's essence, a person's soul, because you cannot change who you are and it is much more difficult to deal with because a hate crime says to a victim of hate speech is that you are not fit to live in society with me, I don't believe that you have the same rights as I do. I believe that you are second to me. I am superior to you.*"

Mandola Legal Analysis
Mr. Hein Dries, Lawyer, Vigilo Consult & Mandola Project

As a good lawyer does if you ask a straight question such as "*What is the definition of Hate Speech?*" you get a very lengthy answer. There are many answers to that question because hate speech has many variants and many legal texts which apply. If you want a full answer Mandola has a very comprehensive document that has many of the legal details and it is recommended that you read it. Now that it is complete, what can be done with it and how can it interact in the project to make sure that it gets used in a way that is meaningful and tells us more than we already know? In the Mandola project there are 10 countries. So that is 10 legal systems, 10 different definitions of what is or what could be hate speech. All the legal texts that could - even remotely - be relevant to answering that question "What is hate speech?" were studied. There are 4 international legal instruments of some authority that not all of the countries are party to. Countries have many particularities which are listed in the document.

The issue to speak to it in simple terms - is that there are many different behaviours that we include so it is really difficult if you ask a lawyer to come up with one single answer. Hate speech can be qualified in legal terms with many different legal articles and we considered visuals aswell. For instance one of the behaviours covered is what if someone's image is used with a text. That is going to be very hard to detect. So what is that definition? Well, it is a BIG mess. From that perspective Europe has a bit of homework and this recommendation is definitely an outcome because the legal norm is not very easily implementable in technical solutions. We have a good analysis of it but actually making this operational is quite a piece of work

For example, 10 countries criminalise incitement to hatred, but 8 of them criminalise incitement to violence. In Ireland they call it "stirring up hatred" – contrary to most countries that speak of "inciting to hatred" and in the first 10 years of existence of the Irish act they never managed to get a case prosecuted, so there is no body of case law. 8 countries additionally criminalise the incitement to discrimination, not mentioned in the framework directive, but party of the convention on cybercrime. 3 countries impose an additional condition about saying something in public – so we may also need to determine what is "public" in the cyber age.

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

From a legal perspective, the Mandola legal work-stream is working to identify what can be easily automated. It works to identify the baseline of hate speech definitions to enable the display of heat-maps. We created three categories. Firstly there are behaviours that are illegal in ALL Mandola's countries. Secondly there are behaviours that are illegal in SOME countries. Thirdly there are the remaining behaviours and special issues that present more challenges.

A significant output is the type of mapping; a type of understanding of how bad it is. Where is society going with it? Is it happening here? Is it happening there? What's happening where? This is the sort of information we are looking to produce and this will influence policy makers, it will influence other people. It is easy to do the automation of negative sentiment or even words related to hatred in an automated fashion. The issue is qualifying it. This is a crucial stage because it will define what is actually on that map or maybe what categories are on that map. It is one thing to know roughly whether people are having negative sentiments in the world, it is another thing if you want to initiate, for example, notice and takedowns with that information.

There is also the issue about what we are actually coding. What is that thing? How transparent is it? The legal work-stream researches this issue since there is a need to review ethics about this activity. What is the data that is in there? Which part of the data was used, not used? The code that we have, quite clearly, is not law. Also, there are further legal issues that need to be addressed including liability and human rights. There is quite a body of law which describes what an ISP should and should not do online for purposes of having human rights and for having freedom of speech.

This area of law is really not easy. If you look at the factors: you have a whole bunch of factors that go into a court decision that decides whether something is or is not hate speech: context, intent, location, medium. Not all of this unfortunately can probably be automated, so Mandola cannot - probably build the full and complete heat map that shows illegal hate speech is in a specific location. The attempt is very worthwhile because we will learn quite a lot about this and about what could be improved to make better, more uniform definitions that would allow this qualification to be more intuitive.

There is a legal regime about hosting providers - the liability regime. This is something not addressed in any of the Mandola deliverables but from an ethical perspective is very important because it will limit what we can ask of Facebook or Microsoft with the law behind us. If an internet service consists of storage or hosting where the content is provided by a third party and the hosting provider does not know about the content then the provider is not liable - provided he does not have actual knowledge of the illegal nature. They only become liable once they have actual knowledge.

Coding such definitions is not easy from a legal perspective.

Session 3
Industry Initiatives

Mr. Andrea Monti, EuroISPA
Industry and the challenge of online hate speech

Hate speech is not absolutely new online. What has changed over time is that the number of internet users has grown significantly many are not accustomed to the online relationship nor the online dialogue. They were given a very powerful tool without any knowledge how to use this tool.

As soon as the Internet was a sort of tool prosecutors and law enforcement received a huge number of crime reports - often petty crimes but crimes nevertheless. There is no single police in the world able to investigate each single claim related to the internet. While you are not aware that a crime has been committed there is no duty of investigation. However, as soon as one is made aware that the crime has been committed, it is mandatory in many jurisdictions, for many prosecutors to start an actual investigation.

If you speak openly with law enforcement they will say that they don't have the resources to deal with hate speech. They are concerned with drugs, arms dealing, terrorism and violent crimes. If those who espouse hate speech are criminals then there are laws that already punish them.

Whatever perspective one has when looking at hate speech, it falls within actual crime that already is punished by the criminal court. You can threaten violence against someone. It is a crime. You can libel him. It is a crime. You can put psychological pressure or you can stalk him. That is a crime. You can invite people to commit another crime. That is a crime in itself. You can associate with other people to commit other crimes. This is a crime in itself.

Why are we so stuck to the importance to the rule of law? When you talk about criminal behaviour remember that every constitution says that there

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

is no private justice. There is only public justice. Otherwise, we turn into a vigilante based society.

For a long time, ISP's have been the target of all the requests for the removal of content or blocking of content and were faced all the time a very schizophrenic attitude of the legislature. This is because on the one hand politicians want to have this social problem sorted out. They did not want a column on a major newspaper that they don't fight such stuff. On the other hand, there is not enough police and prosecutors to deal with this huge number of crimes. The political response has been to just put the responsibility on the ISP shoulder.

The industry has no problem in complying with a court order and does not require a Supreme Court decision that is absolute. A magistrate can assess the situation and issue a seizing order, a blocking order and so on. It is not reasonable to say that "*since the magistrate are understaffed and underprepared then the ISP or the industry should do something*" because ISP's are not the police. If you want industry to do police work and justice work give industry this legal status. Of course it is not possible and industry does not want it. Justice is a public matter. The fact that the government and the state handle the justice is a democratic guarantee.

There is a disturbing trend, not based on reasoned argument but just on practical necessity out of inefficiency, that tends to place on the industry shoulders all the burden of dealing with illegal content, illegal behaviour because the state has no money, no time and has no will to prosecute these issues.

Service providers are companies, they make profits. If someone actually wants that service providers are accountable they can do it. But what about your right to stand for your rights? If some private company based on Mars or Mercury can do whatever it wants? Whatever it wants with your ideas? With your passions? With your beliefs? Do we actually want that the private companies can issue judgements on your thoughts?

ISPAI/www.hotline.ie, Ireland
Mr. Paul Durrant

The Irish Internet Hotline was established in November 1999. It is operated by the ISPAI within the context of a self-regulatory system agreed with Government. Its primary objective is to act against Child Pornography (Child Sexual Abuse Material - CSAM) but it also handles reports of Hate Speech aimed at groups within Ireland. Both constitute illegal content as specified under Irish legislation. CSAM is illegal when it meets criteria as set out in the "Child Trafficking and Pornography Act, 1998 to 2004" and Hate speech is illegal when it meets criteria as set out in the "Prohibition of Incitement to Hatred Act, 1989".

As Chief Executive of the Internet Service Providers Association Ireland (ISPAI) Mr. Durrant has managed the industry-run and government/police supported Internet Hotline for the Republic of Ireland since 2003. He has a technology background, having worked previously in ICT, holding Management, Consultant, Software Engineering and Programming positions.

To be successful a Hotline must be established in such a way that it can be accepted by both the Public and by the Internet industry who must act with confidence on the "notice for takedown" that it serves.

Notices for removal of content should ideally be issued by a Court where the material has been before a judge and therefore ISPs are not liable for actions they take by following that order. If ISPs act without such an order, they may be challenged for wrongful removal of content when the content owner subsequently proves it is not illegal.

It is relatively straight forward and safe for service providers to act against online CSAM, as it is not very likely that paedophiles (who shun publicity) will raise challenges. It is very much the contrary for Hate Speech where the objective of the perpetrators is to have as much publicity of their ideals as possible.

There is so much reported material that may be potentially illegal that as a result the courts do not have enough time for every report to go before a Judge. A compromise to this is the Hotline structure where non-judicial/non-police staff can assess content under the same criteria defined in legislation as a court would use but this is done under some form of self-regulation; co-regulation or government regulation, without the attendant, and usually slow, processes required to bring something before a Court. However, Hotlines if acting somewhat like a court must be accountable and accepted by society.

To be successful a Hotline must provide a trusted, widely known and Government approved service and have accredited assessment staff who dispassionately and accurately appraise reports as to their legality. The Hotline should preferably be established in legislation and have legal authority to issue notices which an ISP can act on without

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

fear of future litigation for having removed content. Due to the global nature of the Internet, National Hotlines have to have international reach. Hotlines should a system to record the reported complaint and all the assessment steps and decisions made by the Hotline staff in deciding why content was found to be legal or illegal.

The Government must support that the Hotline approach is a relevant and effective tool and has a serious role to play in curtailing the distribution of illegal content on the Internet. Law Enforcement, must also support the Hotline, recognising that it is not a competing, but a complementary, resource. The Public must be made aware of the Hotline service but they must be assured of its independence and confidentiality. The Internet Industry must have complete confidence that notices issued by the Hotline are based purely on assessment of the content in line with legislation and not on any moral, political, religious, or other agenda, etc.

There are such differences between countries on what constitutes hate speech that establishing a wide network that can work effectively and swiftly will be very difficult. Language is huge barrier to assessment that essentially is not present when dealing with CSAM images. In additional, extreme hate speech is sometimes promoted by very dangerous and zealous ultra-left or ultra-right groups. These would present a real threat to Hotline staff who they might perceive as trying to suppress their political messages. It cannot be taken as a foregone conclusion that Hotlines in their current form would be equally effective in dealing with Hate Speech.

Mandola/Universidad Autónoma de Madrid
Ms. Paloma Diaz

The Mandola dashboard will visualize large-scale statistics of the spread of potential on-line hate-related speech via Twitter and the Web. In addition to the twitter feed analysis, the sources that are being considered at the moment are comments on articles in online newspapers. The tools that have been developed include spiders which perform web crawling and they extract information from these sources. They extract information about the content, the geolocation of the service and the timestamp .

Mandola is protecting personal information according to the terms of data protection regulations. The processing and storing is in line with article 7(1)(2) of the Personal Data Protection (Protection of the Human) Laws of 2001 to 2012 in Cyprus (N. 138(I)/2001 as it was modified with N. 37(I)/2003 and N105(I)/2012), which was submitted to the Cyprus Data Protection Commissioner on 18th December 2015. Specifically, no sensitive data is stored during the processing. The only information stored is a) the hate processing output, b) the date that it was published or updated, c) the language and d) the location. The location is converted in geo-hash with accuracy reduced to the level of city.

Mandola has recruited social scientists in each country of the Mandola consortium. Their job is to characterize a set of tweets (and other text from the Web) as hate speech, and classify them into individual hate speech categories. The output of this work will be used to create a "ground truth" data set that will be used to train and validate the Mandola classification algorithms.

We would like to demonstrate the dashboard. This is the website based in Cyprus at the moment that we have access to and this is just the early processing of the early processing of some of the tweets and displaying them on a map. If I can pick Ireland for example… it takes a few moments for the updates to arrive from the database…so this is being generated from the database we have. We still have a bit of fine tuning, fixing and correcting to do at this point of time but this is the strategy of what we are trying to achieve and this is the way we are trying to achieve it.

Reports from the front line of Online Hate Speech

Mandola Advisory Board meeting (Oct 2016)
Evangelos Markatos, FORTH,

The first Mandola advisory board took place in Brussels in Early October organised by Dr. Nikos Frydas. We had 15 external and 9 internal members participating. The Advisory Board give us advice to help us improve the project. It is an external point of view. We would like to seek expertise outside Mandola to complement existing threads, counsel on issues raised by Mandola, provide ideas and share experiences. Several members of the advisory board have leading positions in NGO's, academia, and other places so they have a lot of experience that we may not have. The most important and most useful part of the advisory board meeting was the brainstorming sessions. We posed 8 specific questions to the advisory board and we asked them for their views.

We asked participants "What do you think will be the most pressing category of hate speech today?"

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

The result was racism at 32% and migration was second at 27%.

We asked them to think about the future. So think 5 years ahead. So think 5 years down the road. What will be the most pressing issues in hate speech? The first one is racism at 32%, the second one is "others" at 30% which means that they gave a variety of answers. The third was Migration/Refugees at 19%.

Suppose you can pass one law about hate speech. What would that law be? Joint number at 20% was that they would support more freedom-of-speech and they would like a clear simple definition of hate speech. Joint third at 15% was that there was no need for more law and improved internet industry responsibilities.

Are the reporting mechanisms of hate speech today enough and if not, how would you improve them? Are they enough? One answer was the creation of a single report centre led by Europol supported by one report centre in each country country linked to this European centre.

What are the difficulties for industry to respond in this area? Interestingly the first was legality at 27% and the second was freedom-of-speech at 24% because they don't know if something is freedom-of-speech or hate speech, legal/illegal. Third was complexity at 20%.

Are there working models for this space? 39% INHOPE, 21% were not aware of any working models. They were not aware about these reporting mechanisms. They were not familiar about models used by INHOPE, INACH or others. There is a gap about dissemination, awareness. INACH received about 11% and about 7% of people mentioned Europol. It is very surprising that national law enforcement is not on the list – police have been around for 150 years.

What are the challenges for current reporting points responding to hate speech? The first was legal issues at 27%, the second was Reporting/ Analysis systems at 26% and the third was effectiveness at 19%.

The final question was when does hate speech lead to Hate Crime? "Do you see a correlation between hate speech and hate crime?" So if you see a lot of hate speech, do you think you will see a lot of hate crime? Or vice versa? It was an open question. We received a lot of responses and maybe I selected one of them here. "Lack of response to stop it at an early stage leads to a chain of hate speech messages that encourage to go further. So you should stop it in the beginning. If you let it go, it will grow.

The advisory board was actually very exciting and very interesting and very diverse views.

Smile of the Child, Greece
Mr. Marc Van Den Reek

The Smile of the Child is a civil society organisation in Greece that exists since 1996 and the national operator of the 116000 line for missing children, the 116 111 European line with regard to child assistance and then the national 1057 line. The organisation deals with child protection and child care with 4 call centres staffed with professional social workers and psychologists exclusively. There are no volunteers involved. It operates on a 24/7 basis and receives between 250-300,000 calls annually.

This session focusses on the issue of hate speech. There are several other issues faced by the organisation including family violence, abuse, - often sexual abuse on children - and also bullying and hate speech. The three main elements of hate speech encountered are sexism, racism and refugees and very often in the context of issue with regards to missing children.

Bullying is one of the issues but it is very difficult to sometimes determine as a continuum from bullying to hate speech. At a certain point it is difficult to differentiate what hate speech is at a certain moment and what bullying is. When does it become personal and when does it become private and individually oriented or when it is about a group. It is not always easy to make out.

Of course next to the call centres, and the hotline operation, there are the digital platforms. There is a website with a chat room, and a significant focus on Facebook, on Twitter and on other social media and they are targets of hate speech there aswell.

When there is something coming up that is a serious problem with regard to hate speech that is clearly beyond the line of what is illegal the organisation communicates with Facebook and has had very positive experiences with Facebook in that regard. The major part of hate speech that reaches the organisation in the context of the digital platforms is actually YouTube. YouTube is the biggest problem.

Prevention is a very important. On the basis of an MoU with the Ministry of Education in Greece, the organisation has teams of psychologists and social workers (staff also NOT volunteers like in the call centres) who go around to schools and they speak

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

to pupils, to educators and to parents. They also touch upon is hate speech - the collective aspect of targeting a collective group. They touch upon this with students, with pupils, with educators and parents on two basic aspects. The first aspect is sensitisation about the aspect that it is illegal. In many cases people do not know. The second aspect is reporting. They try to enter dialogue with social workers, psychologists, pupils, educators and parents to ensure that this virtual wall between the one who is victim and the one who witnesses hate speech in order to encouraging reporting. In the first half of 2016, 130 schools were reached and about 6,000 pupils, parents and educators.

Smile of the child would very much like to have a clear concise legal definition of hate speech but given the fact that it is not there, it will continue doing the reporting, erasing things from the website, and so on including the grey areas. In principle, although it should stick to what is illegal and nothing else but when you are focussing on a context with children, you do not really have that luxury.

International Cyber Investigation Training Academy
Ms. Maya Boycheva-Manolcheva, Project Manager & PR Expert

ICITA is a non-for-profit organisation NGO, established 7 years ago in Bulgaria. The three main activities include training in cybersecurity and cybercrime investigation, awareness raising and consultancy.

The Mandola strategy for dissemination and awareness it is based on a three pillar approach including traditional dissemination, online dissemination and the external advisory board. Since the start of the project, Mandola has conducted 11 external events, training, conferences, discussions, meetings and reached over 1200 target group representatives, over 20 media participations and 2 press releases. During the first year of project implementation we concentrated mainly on internal communication, but during the second year in which we are already we should concentrate more on the external communication.

The next steps for the Mandola dissemination and awareness strategy is to include more industry, NGO, academic and policy-makers on board and to generate greater interest via external events, by extending geographical media impact and wider social media coverage.

Social activist and former First Lady Eleanor Roosevelt once stated that "*great minds discuss ideas; average minds discuss events; small minds discuss people*". Mandola is open to creative ideas and input from everyone on how to raise the awareness on the problem of online hate speech.

Comments from participants

"Our modern societies would aspire to have a legal system that reflects society's intentions correctly rather than badly representing them."

"When you talk about criminal behaviour remember that every constitution says that there is no private justice. There is only public justice. Otherwise, we turn into a vigilante based society."

"Who is more guilty? A prosecutor that does not follow the rules [of proper data preservation/disclosure protocols] and jeopardises the outcome of a trial or an ISP that asks for court order [and possibly delaying an investigation] to actually deliver legally useable information? "

"The Internet simply moves too fast for the judicial route to be practical as the first level response."

"It cannot be taken as a foregone conclusion that Hotlines in their current form would be equally effective in dealing with Hate Speech"

"there are ways to report either online dangers or hate speech in this case but in my opinion the biggest problem is that it is not simple and not fast or easy sometimes to make a report"

"why don't we all think of instinctively walking into your local police station to complain about a hate crime. Why is that?"

"it is very difficult to sometimes establish it as a continuum as bullying and hate speech in a certain way."

"When does it [bullying] become personal and when does it become private and individually oriented or when it is about a group. It is not always easy to make out."

"When there is something coming up that is a serious problem with regard to hate speech that is clearly beyond the line of what is illegal we do speak with Facebook and we have a positive experience with Facebook in that regard."

"the major part of hate speech that reaches us in the context of the digital platform is actually YouTube. YouTube for us is the biggest problem."

"we would die to have a legal definition of hate speech"

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020

The MANDOLA Project

The MANDOLA project has two main innovations. The first is the extensive use of IT and big data to monitor and report on-line hate speech, and the second is the research on the possibility to make clear distinction between legal and illegal content taking into account the variations between EU Member States legislation.

The project focusses on witnesses of on-line hate speech incidents - who will have the possibility to report hate speech anonymously; policy makers - who will have up-to-date on-line hate speech-related information that can be used to create adequate policy in the field; and ordinary Member States citizens who can have a better understanding of what on-line hate speech is and how it evolves, will be able to recognize legal and illegal on-line hate-speech and will know what to do when they encounter illegal on-line hate.

The MANDOLA project addresses the two major difficulties in dealing with on-line hate speech: lack of reliable data and poor awareness on how to deal with the issue.

Although in general on-line hate speech seems to be on the rise, it is not clear which Member States seem to be suffering most. It is not even clear which kind of on-line hate speech (e.g. homophobia, Xenophobia) is on the rise. Moreover, the available data generally do not distinguish between illegal hate speech and non-illegal hate speech.

Different legal systems in EU Member States make it difficult for ordinary people to easily identify illegal online hate speech. It is difficult for citizens to know how to deal with illegal hate speech and to know how to behave when confronted with harmful but not illegal hate speech. Without reliable data it is very difficult to make reliable decisions and select appropriate policies

In order to achieve the set up objectives the project envisages several activities.

1. Analysis of the legislation of illegal hate-speech at European, international and national level in 10 countries (incl. France, Spain, Greece, Cyprus, Ireland, Bulgaria and the Netherlands, etc.) has been conducted. The legal and ethical framework on privacy, personal data and other fundamental rights protection will be identified and analysed.

2. A monitoring dashboard has been developed which identifies and visualizes cases of on-line hate-related content through social media.

A multi-lingual corpus of hate-related speech has been created based on the collected data. It is used to define queries in order to identify content that may have hate-related speech and to filter the content during the pre-processing phase. The vocabulary has been developed with the support of social scientists and enhanced by the Hatebase (http://www.hatebase.org/).

An API receives the content that may contain hate-related speech and analyses them to determine sentiment scores around keywords, phrases and text. The results and sentiment scores help identify the suspicious content and enhance the hate-related vocabulary. Visualization and advanced reporting approaches will be used in order to present the evolution of on-line hate speech with respect to the time and geographical location of the collected data.

3. A reporting portal allows end users to report potentially illegal hate-related speech materials they have noticed on the Web. The portal provides information based on the reports of the on-line monitoring dashboard. Law enforcement can have access to the portal in order to investigate criminal activities.

4. A smartphone application is under development which will allows reporting of hate-related speech materials accesses on the Web and in social media. The app will be compatible for iOS, Android and Windows mobile devices.

5. A Frequently Asked Questions document has been created and disseminated. This document answers questions like: What is on-line hate speech? What can Internet Providers do? What can users do if they encounter a hateful video, blog, group or receive a hate e-mail or come across a hate-related web site? What can they do if they become target of hate-related comments on-line? How to protect themselves and their children in social networks? The FAQ document will be disseminated via the project portal and the smartphone app.

6. Landscape and gap analysis. Some countries still do not have methods or structures to handle complaints or reports about hate speech. A situation report of current responses to Hate Speech across Europe will be developed and Best Practices Guide for responding to On-line Hate Speech for Internet industry in the European area will be created and disseminated. A comprehensive survey among key stakeholders - major Internet Providers and Law Enforcement is planned. They will identify the key challenges and best practices in responding to hate speech transnationally - especially in countries where freedom of speech is constitutionally protected.

JUSTICE
European Commission
Rights, Equality and Citizenship Programme 2014-2020